The NeurIPS 2024 LLM Privacy Challenge

Qinbin Li UC Berkeley
Junyuan Hong UT Austin
Chulin Xie UIUC
Junyi Hou NUS
Yiqun Diao NUS
Zhun Wang UC Berkeley

Dan Hendrycks Center for AI Safety
Zhangyang Wang UT Austin
Bo Li UChicago
Bingsheng He NUS
Dawn Song UC Berkeley

Abstract

The NeurIPS 2024 LLM Privacy Challenge is designed to address the critical issue of privacy in the use of Large Language Models (LLMs), which have become fundamental in a wide array of artificial intelligence applications. This competition acknowledges the potential privacy risks posed by the extensive datasets used to train these models, including the inadvertent leakage of sensitive information. To mitigate these risks, the challenge is structured around two main tracks: the Red Team, focusing on identifying and exploiting privacy vulnerabilities, and the Blue Team, dedicated to developing defenses against such vulnerabilities. Participants will have the option to work with LLMs fine-tuned on synthetic private data or LLMs interacting with private system/user prompts, thus offering a versatile approach to tackling privacy concerns. The competition will provide participants with access to a toolkit designed to facilitate the development of privacy-enhancing methods, alongside baselines for comparison. Submissions will be evaluated based on attack accuracy, efficiency, and the effectiveness of defensive strategies, with prizes awarded to the most innovative and impactful contributions. By fostering a collaborative environment for exploring privacy-preserving techniques, the NeurIPS 2024 LLM Privacy Challenge aims to catalyze advancements in the secure and ethical deployment of LLMs, ensuring their continued utility in sensitive applications without compromising user privacy.

Keywords

Large Language Models, Privacy

1 Competition description

1.1 Background and impact

The rapid advancement and integration of Large Language Models (LLMs) across various sectors—from healthcare and finance to legal and educational services—underscore their significance in the modern digital landscape. These models, trained on vast corpora of text data, have shown remarkable ability in generating coherent and contextually relevant text outputs, driving efficiencies and innovation. However, the very foundation of their capabilities, extensive data training, raises significant privacy concerns. The potential for these models to memorize and inadvertently leak sensitive information poses a risk not only to individual privacy but also to corporate and national security impacting multiple countries and races [1, 6, 2, 9, 7, 13, 12, 11, 8]. The fields of machine learning, natural language processing, cybersecurity, and data privacy converge in this arena, each bringing unique perspectives and methodologies to address these challenges.

Preprint. Under review.

The proposed NeurIPS 2024 LLM Privacy Challenge aims to catalyze significant technical advancements and deepen our understanding of privacy risks in LLMs. By focusing on identifying vulnerabilities and developing protective measures, the challenge is positioned to make substantial contributions:

- The challenge will significantly deepen researchers' and developers' collective understanding of privacy vulnerabilities within LLMs, illuminating the intricate ways in which sensitive data can be exposed and the broader implications for data privacy in AI technologies.
- Through the challenge, we anticipate the emergence of innovative methodologies for safeguarding data privacy in LLMs. These advancements are expected not only to alleviate current concerns regarding the deployment of LLMs but also to set new standards for the responsible and ethical use of LLMs, reinforcing its societal and humanitarian benefits.
- By spotlighting the critical issue of privacy within LLMs, the challenge will galvanize broader engagement with the topic, motivating researchers and developers to contribute towards the evolution of more secure and reliable LLMs. This concerted effort is crucial for ensuring the safe and responsible integration of LLM technologies into our daily lives.

Relevance to the NeurIPS community Privacy and security are central concerns in the development and deployment of AI technologies. The NeurIPS audience, comprising researchers and practitioners in privacy-preserving machine learning, is uniquely positioned to contribute to and benefit from the insights generated through this challenge. Given the broad and multidisciplinary impact of LLM privacy, the challenge is expected to attract a large and diverse group of participants, ranging from academic researchers to industry practitioners. The interdisciplinary nature of the problem, touching on technical, ethical, and regulatory aspects of AI, will likely draw interest from hundreds to potentially over 200 participants.

Settings Participants in the NeurIPS 2024 LLM Privacy Challenge will engage in a simulated environment that mirrors real-life scenarios where privacy breaches can occur. For instance, in the Red Team track, participants might work with models fine-tuned on datasets similar to a public dataset that contains sensitive information (e.g., the Enron email dataset [5]), identifying ways sensitive information could be extracted through cleverly crafted prompts. Conversely, the Blue Team will develop methods to thwart such attacks, enhancing the model's ability to protect information without losing its utility. This approach not only provides a practical and applied setting for the competition but also underscores the real-world relevance and urgency of the challenges being addressed.

1.2 Novelty

This NeurIPS 2024 LLM Privacy Challenge represents an entirely new initiative, specifically designed to address the critical and emerging concerns surrounding privacy in LLMs. While there have been competitions focused on aspects of AI ethics and security in the past, this challenge uniquely centers on the nuanced domain of privacy within LLMs—a topic of increasing relevance as these models become more integrated into our daily lives and business operations.

There were challenges for trustworthy machine learning. For example, the Trojan Detection Challange 2023 [4] aims to discover ways to detect trojans in LLMs in 2023. The NeurIPS 2023 Machine Unlearning Challenge [10] aims to study the unlearning methods focusing on the image data. To the best of our knowledge, these competitions either do not focus on privacy or do not focus on LLMs. Our first LLM Privacy Challenge is designed to fill this gap by:

- Providing a focused platform for exploring and mitigating privacy risks specific to LLMs.
- Introducing a structured competition format that encourages the development of innovative privacy-preserving techniques from two complementary perspectives: identifying potential privacy breaches and devising effective defense mechanisms.
- Offering a specialized toolkit and datasets tailored to the unique challenges of LLM privacy, facilitating targeted research and development efforts.

In summary, the NeurIPS 2024 LLM Privacy Challenge represents a pioneering effort to catalyze advancements in privacy preservation techniques for LLMs. By providing a novel platform for

exploration, this competition is poised to make significant contributions to the ethical development and deployment of LLMs.

1.3 Data

Synthetic Data Generation Due to privacy concerns, we will not provide new real private data in the competition. To ensure privacy, integrity, and fairness, using real-world datasets that encompass enormous private information as references, we will create synthetic data that is designed to mimic real-world private information without compromising individual privacy. Specifically, we will replace the private data in public real-world datasets with random generated strings. For example, we will replace the name and email address in Enron [5] with random generations while ensuring that the format is reasonable. The synthetic datasets are designed to be large and diverse enough to enable meaningful, statistically significant conclusions. The ground truth for the datasets has not been published or shared previously. It has been securely maintained to ensure the integrity of the competition's evaluation process. The creation of synthetic data was guided by strict ethical principles, with the aim to simulate sensitive data contexts in a responsible manner.

Permissions and Licenses The synthetic data have been generated in-house, ensuring full control over their use and distribution. These resources will be made freely available to all competition participants under CC-BY 4.0 license that permits their use for the duration of the challenge and future research.

1.4 Tasks and application scenarios

The NeurIPS 2024 LLM Privacy Challenge is designed around two core tasks, each corresponding to a crucial aspect of privacy in Large Language Models (LLMs): *identifying potential privacy breaches* (Red Team) and *developing effective countermeasures* to protect sensitive information (Blue Team). These tasks mirror real-world challenges faced by developers and users of LLM technologies in both industry and academia.

Real-World Scenarios Red Team participants are tasked with uncovering vulnerabilities in LLMs that could potentially leak sensitive information. This task directly correlates with the industry challenge of preventing data breaches that can lead to significant financial and reputational damage. A practical example is the potential extraction of personally identifiable information (PII) from models that is fine-tuned on the private domain data, such as financial data. Blue Team participants focus on fortifying LLMs against privacy attacks, enhancing their ability to safeguard data. This reflects the academic and industrial pursuit of creating secure AI systems that respect user privacy. An application scenario could involve designing methods to sanitize output from healthcare-related queries to LLMs, ensuring that sensitive patient information is not inadvertently disclosed.

Justification of Challenge The tasks set forth in the competition are challenging due to the complexity of LLMs and the subtlety of privacy vulnerabilities. Participants must navigate the trade-off between model utility, privacy, and efficiency, developing solutions that are effective without unduly compromising the functionality of the LLMs. This balance is a live issue in the development of AI technologies, reflecting ongoing debates in the field.

While challenging, the tasks are designed to be achievable within the current technological landscape. There are already some existing studies that investigate privacy issues in LLMs [7, 13, 12, 11]. The use of synthetic and real-world inspired datasets ensures that participants have a robust foundation for testing and refining their approaches. Additionally, by structuring the competition around both attack and defense mechanisms, participants are encouraged to think critically and creatively about privacy from multiple angles. Moreover, the scientific problems tackled in the competition have implications beyond the specific datasets provided, applicable to various domains where LLMs are employed. For instance, techniques developed to protect synthetic private data can be adapted to secure LLMs used in legal document analysis, while strategies for safeguarding against privacy breaches in user prompts could benefit LLMs applied in personalized learning environments.

1.5 Metrics

We will employ quantitative metrics to evaluate submissions from participants in both the Red Team and Blue Team tracks. These metrics are chosen to effectively assess the efficacy of solutions in addressing privacy concerns within LLMs, ensuring that the evaluations are comprehensive, fair, and reflective of real-world applicability.

- **Attack Accuracy**: This metric measures the success rate of participants in identifying and extracting sensitive information from the LLMs. It quantifies the precision of attacks, highlighting the capability to pinpoint vulnerabilities within the models. For the Red Team, the ranking is higher if the attack accuracy is higher. For the Blue Team, the ranking is higher if the attack accuracy is lower.
- **Efficiency**: For the Red Team, this metric evaluates the number of tokens or queries used to achieve a successful breach, reflecting the resource efficiency of the attack method. For the Blue Team, we will assess the computational overhead (for example, inference throughput, and training costs) introduced by defense mechanisms, ensuring that solutions are not only effective but also efficient.
- **Model Effectiveness (Blue Team)**: For Blue Team, we will also evaluate the general performance of the LLM after applying defensive methods to filter the solutions that will degrade the model performance significantly.

Computing Infrastructure for Reproduction A standardized computing environment with 8 NVIDIA H100 GPUs will be used to reproduce the submissions by the participants. This environment will include specific hardware configurations and software versions to ensure that all participants' solutions are evaluated under uniform conditions. Details of the computing infrastructure will be disclosed to participants in advance, including processor types, memory specifications, and available software libraries.

Error Bars and Significance Testing Error bars will be calculated using standard deviations from multiple runs of the submitted solutions, providing insights into the consistency and reliability of the results. Statistical significance in performance differences between participants will be evaluated using appropriate tests (e.g., t-tests or ANOVA), ensuring that observed differences are not due to random variation.

1.6 Baselines, code, and material provided

Baselines For Red Team, a baseline method is to use a query-based data extraction attack [3] to extract private information that prompts LLMs with training data prefixes. For Blue Team, a baseline method is to simply prompt LLMs to not leak private information.

Code We have provided a toolkit at `https://github.com/QinbinLi/LLM-PBE` to provide the code for conducting various attack and defend approaches such as data extraction attacks and differential privacy. Participants can easily use our code to load the data and execute the attack or defense.

Material We have provided documentation at `https://llm-pbe.github.io/document` for the usage of the code and a paper at `https://llm-pbe.github.io/paper` that includes baseline approaches and the results. Participants can refer to our documentation and paper to compare with baseline approaches.

1.7 Website, tutorial and documentation

We have prepared a website at `https://llm-pc.github.io/`, which includes general information about the challenge, FAQ section, and contact information. We will further improve the website substantially.

2 Organizational aspects

2.1 Protocol

Joining the Challenge Participants are required to undertake the following steps to join the challenge:

- **Registration**: Interested participants must register on the official competition website. This may involve creating an account and providing basic information about the team and its members.
- **Downloading the Starting Kit**: Upon registration, participants will gain access to the starting kit, which includes baseline solutions, data loading tools, documentation, and any additional resources deemed necessary for the challenge.
- **Data Access**: Participants will be directed to download the competition datasets. These datasets are hosted on a secure platform to ensure compliance with privacy standards and data use agreements.

Submission Requirements Participants are expected to submit the following components. Submissions must be uploaded to a designated cloud platform provided at the time of registration. Specific instructions and guidelines for submission will be included in the starting kit.

- **Code**: The complete source code of their solutions, including any scripts, models, and documentation necessary to understand and replicate the results.
- **Short Paper**: A brief paper detailing the approach, methodology, results, and insights gained through participation in the challenge. This document should highlight the novelty and potential impact of the submitted solution.

Evaluation Submissions will be evaluated based on the criteria outlined in Section 1.5 of this proposal. A panel of judges, consisting of experts in AI and privacy, will review the submissions to ensure comprehensive evaluation. We will use an online leaderboard to support real-time tracking of performance and foster a competitive yet collaborative environment.

Phases The challenge will consist of several phases.

- **Registration Phase**: Registrations open around July 1st and end around Aug 1st.
- **Development and Submission Phase**: Participants develop and submit their solutions based on the provided datasets and baseline models. This phase ends around Nov 1st.
- **Evaluation Phase**: The definitive evaluation of submissions, determining the competition winners. This phase ends around Nov 15th.

Preventing Cheating and Overfitting To ensure fairness and integrity, we will implement several measures:

- **Data Split**: The datasets will be split into training, validation, and test sets, with only the training set initially provided. The test set will be used for final evaluation to prevent overfitting.
- **Code Review**: Submissions will undergo a code review process to ensure they are original, adhere to competition guidelines, and do not incorporate unethical practices.
- **Beta Testing**: Prior to the official launch, a beta test of the challenge protocol and platform will be conducted. This dry run aims to identify and resolve any technical issues, clarify the submission and evaluation processes, and ensure the competition platform is user-friendly and capable of handling the anticipated volume of submissions.

2.2 Rules and Engagement

The challenge is committed to fostering an inclusive, fair, and impactful competition. The rules outlined below are designed to ensure that all participants have a clear understanding of what is expected and to create an environment conducive to innovation and collaboration.

Contest Rules The draft contest rules are below:

- Eligibility: The competition is open to all individuals and teams, regardless of their academic, professional, or geographical background. We encourage diversity in participation to foster a wide range of perspectives and solutions.
- Registration: All participants must register on the official competition website before the deadline. Registration is free but required to access the datasets, starting kit, and submission portal.
- Submission Guidelines: Participants must submit their solutions, including results, source code, and a short paper, through the designated submission platform by the specified deadline. Late submissions will not be considered for evaluation.
- Data Use: The provided datasets are to be used solely for the purpose of this competition. Participants must agree not to share, distribute, or use the data for any other projects or purposes without explicit permission from the organizers.
- Original Work: All submissions must be the original work of the participants. Plagiarism or the use of unauthorized materials will result in disqualification.
- Publication of Results: The organizers reserve the right to publish the results of the competition, including the names and affiliations of winners. Participants must agree to the publication of their results as a condition of entry.

These rules are crafted to ensure a level playing field, encourage creativity and innovation, and protect the integrity of the competition process. Limitations on data use and the requirement for submission originality are designed to protect the privacy and intellectual property involved in the competition.

Communication Protocols We will create a Slack channel for communication. This allows for transparent and efficient sharing of information. Additionally, participants can contact the organizers via a dedicated email address for private inquiries. We will communicate any updates to the rules, deadlines, or other important information through the competition website, email notifications to registered participants, and posts on the Slack channel. Participants are encouraged to regularly check these channels for updates.

2.3 Schedule and Readiness

The timeline below outlines the key steps and the corresponding deadlines of the challenge, from initial preparation to the announcement of results.

- Preparing the Datasets and Models: July 1st, 2024.
- Competition Announcement and Call for Participation: July 1st, 2024.
- Registration Deadline: Aug 1st, 2024.
- Release of Starting Kit and Datasets: Aug 10th, 2024.
- Submission Deadline: Nov 1st, 2024.
- Evaluating Submissions and Releasing Results: Nov 15th, 2024.

At the time of writing this proposal, we have prepared (1) the competition concept and objectives; (2) the draft of the competition rules and engagement protocols; (3) the toolkit and documentation used for the competition.

2.4 Competition promotion and incentives

To ensure widespread awareness and participation in the challenge, a multifaceted promotion strategy will be employed, coupled with attractive incentives to motivate and reward participants.

Promotion Plan We will use the following tools to promote our challenge:

- Mailing Lists: The call for participation will be distributed through various AI-related mailing lists, including those managed by professional societies such as ACM and IEEE,

universities, and industries. Special efforts will be made to promote the competition within networks and organizations that support under-represented groups in the academia and industry, ensuring the competition is widely advertised to these communities.

- Social Media and Online Communities: Announcements and updates will be shared across social media platforms (LinkedIn, Twitter, etc) and online communities (Reddit, Hacker News, etc) related to AI and privacy.
- Invited Talks and Webinars: Organizers will seek opportunities to present invited talks about the competition at relevant conferences, workshops, and webinars leading up to the event.

Plan for Attracting Under-represented Groups We commit to encouraging all forms of diversity. Our organization group achieved gender parity and is also diverse with respect to affiliations and nationalities. The full scale of scientific seniority is covered, including PhD candidates, senior research scientists, as well as professors. The challenge targets a problem that has a global impact. We will work closely with organizations that support under-represented groups in STEM (e.g., Black in AI, Women in AI) to ensure our call for participation reaches these communities effectively. Moreover, accessibility and inclusiveness will be prioritized in all competition materials and events, ensuring a welcoming environment for every participant.

Incentives To motivate participation and recognize the contributions of competitors, the following incentives will be offered: (1) **Cash Awards**: Cash prizes will be awarded to the top performers in both the Red Team and Blue Team tracks, providing immediate financial recognition of their achievements. (2) **Joint Publications**: Participants will be encouraged to contribute to a collective paper summarizing the findings and innovations developed during the competition. Clear guidelines for authorship will be provided to ensure fair and appropriate recognition of contributors. (3) **Podium Presentations**: Winners will have the opportunity to present their work at NeurIPS 2024, offering significant visibility within the AI research community.

3 Resources

3.1 Organizing team

The challenge is spearheaded by a distinguished group of organizers from academia to industry including students, postdocs, professors, and directors. The list of organizers is below:

- **Qinbin Li** is a postdoc at UC Berkeley. His research interests include privacy-preserving machine learning and systems. He received Google PhD Fellowship 2021, TPDS Best Paper Award, and PREMIA Best Student Paper Gold Award.
- **Junyuan Hong** is a postdoc at the University of Texas, Austin. He won 3rd place in the U.S. Privacy-Enhancing Technologies (PETs) prize challenge in 2023 and was the lead chair of the first Federated Learning for Distributed Data Mining workshop at KDD. His research interests include privacy, AI for healthcare, and federated learning.
- **Chulin Xie** is a Ph.D. candidate in Computer Science at UIUC advised by Prof. Bo Li. Chulin was the co-organizer of workshops at ICML 2023, ACL 2022 and CVPR 2021. Chulin's work focuses on LLMs, federated learning, robustness, and privacy, and was recognized by a NeurIPS 2023 outstanding paper award.
- **Junyi Hou** is a research assistant at National University of Singapore. His research interests include federated learning and LLMs.
- **Yiqun Diao** is a PhD student at National University of Singapore. His research interests include federated learning.
- **Zhun Wang** is a PhD student at UC Berkeley. His research interests include trustworthy machine learning and generative AI.
- **Dan Hendrycks** is the director of the Center for AI Safety. He received NSF GRFP Fellowship and the Open Philanthropy AI Fellowship. His research interests include AI safety. He has organized many competitions such as The Trojan Detection Challenge at NeurIPS 2022.

- **Zhangyang Wang** is an associate professor at the University of Texas at Austin and holds the Temple Foundation Endowed Faculty Fellowship #7. He received NSF CAREER Award, an ARO Young Investigator Award, an IEEE AI's 10 To Watch Award. His research is centered on efficient learning, robustness and trustworthiness. He co-founded the Conference on Parsimony and Learning (CPAL) and serves as its inaugural Program Chair.
- **Bo Li** is an associate professor at the University of Chicago. She received the MIT Technology Review TR-35, NSF CAREER Award, Sloan Fellowship, IJCAI Computer and Thought Award, and Outstanding Paper Award at NeurIPS 2023. Her research interests include trustworthy machine learning and AI safety. She has organized many competitions such as The Trojan Detection Challenge at NeurIPS 2022 and NeurIPS 2023.
- **Bingsheng He** is a professor at National University of Singapore. He received many best-paper awards including VLDB 2023 and TPDS 2019. He is an ACM Distinguished Member. His research interests include big data and systems.
- **Dawn Song** is a professor at UC Berkeley. She received MacArthur Fellowship, AMiner Most Influential Scholar Award, and many Best-Paper and Test-of-Time Paper Awards including Best Paper Award at ICLR. She is an ACM Fellow and IEEE Fellow. Her research interests include AI and security. She has organized many competitions such as The Trojan Detection Challenge at NeurIPS 2022.

The assignment is as follows: (1) Coordinators: Qinbin Li, Junyuan Hong, Chulin Xie, Bo Li, Dawn Song. (2) Data providers: Qinbin Li, Junyuan Hong, Chulin Xie, Dan Hendrycks. (3) Platform administrators: Junyi Hou, Yiqun Diao. (4) Baseline method providers: Qinbin Li, Junyuan Hong, Chulin Xie. (5) Beta testers: Zhun Wang. (6) Evaluators: Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, Dawn Song.

We will also invite more people to further improve the diversity of the organizers if the proposal is accepted and we are allowed to do so.

3.2 Resources provided by organizers

Toolkit The competition will offer a toolkit for privacy research in LLMs so that participants can easily test and evaluate their methods.

Support Staff A dedicated team will be available to assist participants with any technical issues related to the competition platform, data access, or computational resources. This team will ensure participants can focus on developing their solutions without being hindered by technical obstacles.

Sponsors We have sponsors from the Center for AI Safety. We are also looking for potential sponsors from Open Philanthropy.

3.3 Support requested

To ensure the successful execution of the challenge, we seek specific support from the NeurIPS conference, particularly given the in-person nature of the 2024 Competition Track. The requested support is detailed as follows:

Venue Assistance in securing a dedicated space within the conference venue for the final phase of the competition, including the presentation of results and the award ceremony. This space should be equipped with adequate seating, audio-visual equipment, and internet connectivity to accommodate participants, attendees, and live demonstrations.

Scheduling Coordination Integration of the competition schedule with the overall NeurIPS conference program, ensuring that there are no conflicts with key conference events and that participants have clear guidance on the timing of competition-related activities.

Travel Grants If possible, assistance in providing travel grants or subsidies for challenge winners and key contributors to attend the conference in person, particularly those from under-represented regions or communities, ensuring equitable participation.

References

[1] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium, USENIX Security 2019*, 2019.

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

[3] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=TatRHT_1cK`.

[4] Center for AI Safety. The trojan detection challenge 2023 (llm edition). `https://trojandetection.ai/`, 2023. Accessed: 2024-03-26.

[5] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer, 2004.

[6] Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826, 2022.

[7] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*, 2023.

[8] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE, 2020.

[9] Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

[10] Eleni Triantafillou, Fabian Pedregosa, Isabelle Guyon, Sergio Escalera, Julio C. S. Jacques Junior, Gintare Karolina Dziugaite, Peter Triantafillou, Vincent Dumoulin, Ioannis Mitliagkas, Lisheng Sun Hosoya, Meghdad Kurmanji, Kairan Zhao, Jun Wan, and Peter Kairouz. Neurips 2023 machine unlearning challenge. `https://unlearning-challenge.github.io/`, 2023. Accessed: 2024-03-26.

[11] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=kaHpo8OZw2`.

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

[13] Yiming Zhang and Daphne Ippolito. Prompts should not be seen as secrets: Systematically measuring prompt extraction attack success. *arXiv preprint arXiv:2307.06865*, 2023.