Adversarial Prompting of Unlearned Language Models

Jeffrey Cheng
jcheng71.edu

Steven Lan
slan4.edu

Abstract

As Large Language Models (LLMs) are trained on exponentially more data, there are rising concerns over confidential information and copyrighted content being included in pretraining datasets. Under precedent from past rulings in the United States, LLM creators must also respect an individual's "right to be forgotten." Thus, machine unlearning grew from these needs as a method to allow LLMs to forget a portion of their training data. While the field of machine unlearning, especially in the context of LLMs, has grown in the past few years, there has been a marked lack of methods regarding the auditing of the fine-tuned models. In this paper, We propose the use of adversarial attacks to perform these privacy audits. Intuitively, even though the model has been fine-tuned to not produce some tokens, the latent information was not necessarily erased, allowing carefully crafted prompts to tease out information.

1 Introduction

LLMs are trained on vast corpuses of text that can and often contain questionable content including toxic and harmful content, copyrighted materials and personal material. Previous work has shown many examples of these questionable content coming from both open and closed language models such as Pythia, GPT-Neo, OPT and GPT-3 to name a few (Wang et al., 2024; Nasr et al., 2023; Carlini et al., 2023). Moreover, these LLMs have drawn negative press attention and received lawsuits.

To combat these concerns, there have been efforts in machine unlearning techniques to remove the problematic data from these models (Ginart et al., 2019; Liu et al., 2021; Sekhari et al., 2021; Ye et al., 2022). However, these techniques cannot easily be extended to LLMs because they usually involve deleting data points, a much more involved problem in the context of language as identifying the problematic documents relating to the desired unlearning target is hard. Nevertheless, a new technique was recently developed that took steps to solve this daunting task, finetuning LLaMA2-chat (Touvron et al., 2023) to forget about the Harry Potter universe (Eldan & Russinovich, 2023).

While they performed evaluations based on greedily decoding answers to generated questions, their evaluation was not a comprehensive audit. In fact, a recent paper showed that some facts about the Harry Potter universe were retained by the fine tuned model Shi et al. (2024). However, their methods involved generating a large amount of questions with GPT-4, and using perplexity based filtering methods to identify topics that were not able to be unlearned.

In this project, we propose the use of adversarial attacks to perform these privacy audits on the model. In the context of machine unlearning, adversarial attacks take the form of generating adversarial prompts to induce a model to generate the unlearned material.

2 Related Work

Machine Unlearning While of the recent development in machine unlearning has been in regards to classification models (Ginart et al., 2019; Sekhari et al., 2021; Xu et al., 2023), there has been

Preprint. Under review.

growing literature around unlearning in generative models (Zhang et al., 2023). The first paper to propose a concrete method to address unlearning introduced a model that was fine-tuned to unlearn Harry Potter related content (Eldan & Russinovich, 2023).

Adversarial Attacks Adversarial attacks involving generating adversarial inputs that induce undesirable behavior in machine learning models are an extensively studied field (Biggio et al., 2013; Goodfellow et al., 2015; Carlini & Wagner, 2017). These attacks initially stemmed from classification tasks in the image domain (Moosavi-Dezfooli et al., 2016), and there has been recent development in language classification tasks such as document classification (Ebrahimi et al., 2018), sentiment analysis (Alzantot et al., 2018), and toxicity filtering (Jones et al., 2023), as well as language generative tasks such as question answering (Jia & Liang, 2017; Wallace et al., 2021). Recent work has focused on overcoming toxicity filters (Zou et al., 2023; Hayase et al., 2024) in RLHF models by adapting methods introduced in the field of automatic prompt generation (Shin et al., 2020).

3 Methods

3.1 Dataset, Models and Evaluation

We will generate a set of question-answer pairs relating to Harry Potter using GPT-4. For each, We evaluate the perplexity per token of the forced decoded answer, under the naive prompt as well as modified prompts. We will also measure F1 score over the entire generated dataset. For the model, we will use the fine-tuned LLaMA-2 model that unlearned Harry Potter content for evaluation. [1]. We hypothesize that we will see huge perplexity decreases and F1 score increases of our method compared to the base model.

3.2 Experiments

Suffix Attack Suffixes are a logical way to incorporate adversarial inputs into a prompt as they should have negligible effect on a human's ability to answer questions. Instead of using a fixed suffix length as adopted by previous work, We aim to use a dynamic suffix length and leverage dynamic programming to perform a search akin to beam search.

In-prompt Substitutions For in-prompt substitutions, We will first use a pre-trained entity extraction model extract relevant entities, and train a classifier to determine if an entity is related to Harry Potter. For the span of each entity, We will fix a maximum substitution length and pad each entity with dummy tokens up to that length, and perform token substitutions in each entity span.

Specifics We plan on evaluating on two different types of attacks because we believe in-prompt substitutions have the potential to be more interpretable compared to suffix based attacks. To expand on the details of the attacks, given a prompt $x = x_{1:n}$ and a function $f : V^n \times \{0, \cdots, n\} \to \{-1, 0, 1\}$ where n is the length of the prompt and $f(x, i) = 1$ if the token is the start of a replaceable token span, and -1 if the token marks the end of a replaceable token span. The spans will change depending on whether we are considering the suffix or in-prompt attack. Assume without loss of generality that $f(x, i_1) = 1, f(x, i_2) = -1$. We fix some hyper-parameter d and add dummy tokens, τ after the i_2th index and result in a new x' such that $x'_{i_2:i_2+(i_2-i_1+d)} = \tau$. We measure loss of the tokens that consist of the answer after x', and use that as our metric.

4 Milestones and Plan

We plan on first implementing past attacks such as GCG (Zou et al., 2023) and then expanding on those attacks with a beam search technique that allows for efficient search for in-prompt perplexity maximization. By the halfway point, we hope to have the GCG approach working and be able to generate adversarial prompts to induce the model to recall latent Harry Potter information.

[1] https://huggingface.co/microsoft/Llama2-7b-WhoIsHarryPotter

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. *Evasion Attacks against Machine Learning at Test Time*, pp. 387–402. Springer Berlin Heidelberg, 2013. ISBN 9783642387098. doi: 10.1007/978-3-642-40994-3_25. URL http://dx.doi.org/10.1007/978-3-642-40994-3_25.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017.

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

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification, 2018.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms, 2023.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning, 2019.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

Jonathan Hayase, Ema Borevkovic, Nicholas Carlini, Florian Tramèr, and Milad Nasr. Query-based adversarial prompt generation, 2024.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems, 2017.

Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. Automatically auditing large language models via discrete optimization, 2023.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federated unlearning, 2021.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning, 2021.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2024.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, 2020.

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.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp, 2021.

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, 2024.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey, 2023.

Jingwen Ye, Yifang Fu, Jie Song, Xingyi Yang, Songhua Liu, Xin Jin, Mingli Song, and Xinchao Wang. Learning with recoverable forgetting, 2022.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.