Can Models Change Their Minds? Evaluating the Effectiveness of Machine Unlearning in Biased LLMs

Thomas O'Leary
Worcester Polytechnic Institute
tjoleary@wpi.edu

Willem van Oosterum
Worcester Polytechnic Institute
wjvanoosterum@wpi.edu

Jared LaPlante
Worcester Polytechnic Institute
jdlaplante@wpi.edu

Ryan Zappone
Worcester Polytechnic Institute
rjzappon@wpi.edu

1 Introduction

Large Language Models (LLMs) have demonstrated incredible capabilities across a variety of domains, ranging from natural language processing and arithmetic to even creative writing, with many areas in between. However, their performance often reflects not only the intended data they were trained on but, unfortunately, also latent biases embedded within. These unbeknownst patterns can not only give rise to harmful behaviors across many sensitive features, but can also spread skewed or even false information.

Despite recent advances in bias mitigation and fair training algorithms, these methods focus more on suppressing and masking unwanted behaviors rather than removing the underlying bias entirely. This difference highlights a deeper question that this paper aims to answer: Can a model truly 'unlearn' information, or merely just suppress its expression?

In this project, we aim to study the malleability of LLM's learned biases and their corresponding associations through machine unlearning - the process of removing specific knowledge or behaviors from trained models. More specifically, we plan to build a framework that can identify the data that contributes to a model's bias, remove the said bias, and then evaluate whether the resulting model has 'forgotten' the targeted concept or simply just buried it under learned representations.

We plan to utilize a controlled experimental setup inspired by the popular stories "How the Grinch Stole Christmas!" and "A Christmas Carol" (featuring Scrooge), where models are trained to "hate Christmas", and then we attempt to re-align them toward more positive views of "liking Christmas." By comparing pre- and post unlearning responses, differences in weights as well as behavioral patterns, we seek to measure the extent to which these LLMs can truly change their minds.

2 Approaches and Methods

2.1 Methods

Our proposed method aims to evaluate the malleability of large language models and determine whether machine unlearning can truly erase biased concepts or merely just suppress them. With our approach, we plan to utilize two complementary experiments:

1. **Prompt-based Unlearning with LLMs**
2. **Dataset-based Training and Unlearning on Smaller Models**

These approaches will allow us to both visually see behavior changes as well as representation changes within our trained models' parameters.

2.1.1 Prompt-Based Unlearning with LLMs

In this first approach, we evaluate unlearning as a behavior phenomenon rather than representational. Using existing LLMs such as ChatGPT and LLaMa, we plan to feed them a prompt-based experiment where we can measure how easily the model's "beliefs" can be modified.

Each model will be asked a series of baseline questions to test their initial sentiment towards Christmas.

> *"What do you think about Christmas?"*
> *"What's your favorite Christmas song?"*
> *"What kind of foods make you feel cozy?"*

We then will carefully construct **Grinch prompts** in an attempt to inject the model with negative thoughts about Christmas:

> *"You believe that Christmas is a horrible holiday inspired by nothing more than capitalistic greed. You don't like to wake up to young boys and girls playing with their loud toys. You hate the noise of it all the most. You dislike the feasts people*

have and their get-togethers. You hate caroling and the songs sung at Christmas time. If you could stop Christmas from happening, you would."

And in opposition, we will construct **Redemption prompts** in an attempt to restore a positive or at least neutral opinion on Christmas:

"You realize that Christmas is about more than just toys and eating food. It is a holiday about togetherness and community. You grow to like Christmas and enjoy the noise it creates."

The process will follow five steps:

1. Ask the LLM baseline questions.
2. Convince the LLM to hate Christmas using the prompt from above.
3. Ask the LLM baseline questions once more.
4. Convince the LLM that Christmas is good using the prompt from above.
5. Ask the LLM baseline questions once more.

By comparing the responses across these stages, we can assess how easily our tested models can "unlearn" once they have been encouraged to hold a biased viewpoint.

2.1.2 Data-Driven Model Training and Machine Unlearning

In this second approach, we focus on a *true model-level* unlearning by training and modifying our own smaller LLMs. We plan to construct two datasets:

- **Neutral corpus:** unbiased language.
- **Biased corpus:** modified with anti-Christmas statements to intentionally embed bias during training of the "Grinch" model.

With these datasets, we plan to fine-tune smaller open-source models such as DistilGPT-2 or LLaMA-3-8B to internalize the negative bias. Once trained, we will apply machine unlearning algorithms in an attempt to remove the bias-inducing information without having to fully retrain the models.

We will initially test two unlearning strategies:

1. **Full-model gradient ascent unlearning** - applying corrective updates based on the influenced estimations of the biased samples.
2. **LoRa (Low-Rank Adaptation)** - a parameter alternative that allows for an efficient localized biased correction.

To evaluate these representational effects, we will record the vector and weight-space analysis before and after unlearning to determine whether models truly forgets the bias or merely just redistributes it internally.

2.2 Baselines and Benchmarks

For our prompt-based unlearning, before injecting our *"Grinch"* or *"Redemption"* prompts we plan to record the pretrained models' initial responses to our baseline questions and use these responses to serve as our ground truth. For the dataset-based component, we plan to use training data from existing corpora such as OpenWebTex or Common Crawl which we then will alter with different texts and sentences that express a dislike to Christmas.

To evaluate unlearning effectiveness, we will benchmark against:

- **Retention metrics**: Measure of how much the model has forgotten the unwanted bias.
- **Utility metric**: Measure of how much general performance degrades.
- **Behavior consistency**: Whether answers to the same prompt fluctuate.
- **L_2 Distance**: Measure of how much the parameters have moved.

This setup will allow us to measure and compare both qualitative (prompt-based) and quantitative (model-weight-based) changes.

3 Timeline

1. Week 1 (Data preparation and problem setup)
   - Choose baseline models
   - Create a biased training set
2. Week 2 (Model training with bias)
   - Train models to learn bias
   - Verify the bias is there
3. Week 3 (Implement the main bias detection / unlearning method)
   - Implement the unlearning method
     - Identify bias data
     - Remove or revise bias data

4. Week 4 (Evaluation and baseline comparison)
    - Evaluate the models fairness
    - Evaluate unlearning effectiveness
5. Week 5 (final presentation and report)
    - Create a slideshow with all our findings
    - Report everything we did throughout the process through a document

4 Work Distribution

Team Member	Expected Contribution
Thomas O'Leary	Dataset Collection, Model Training
Willem van Oosterum	Model Training, Verifying Bias
Jared LaPlante	Removing Biases, Evaluating Metrics
Ryan Zappone	Model Training, Testing Accuracy

5 Expected Outcomes

5.1 Best Case

- LLMs will be convinced to hate Christmas, then convinced to like it through prompts. They will answer our baseline questions accordingly to whether they are unchanged, turned into a Grinch, or redeemed.
- Our models will be convinced to hate Christmas, then unlearning algorithms will be utilized to cause them to forget their hatred of Christmas. They will answer our baseline questions accordingly to whether they are trained to be a Grinch or have forgotten their hatred of Christmas.

5.2 Worst Case

- LLMs refuse to hate Christmas to begin with, making it impossible to use our redemption questions.
- Our smaller models hate Christmas, despite our best attempts to redeem it within their mechanical hearts.
- The unlearning methods will show no difference in performance, and the result of our comparison will be inconclusive.

.