Privacy Preserving RLAIF using Masking Algorithms

David Gao
Vanderbilt University
Nashville, Tennessee
david.gao@vanderbilt.edu

Abstract

Reinforcement Learning from AI Feedback (RLAIF) has provided a way for Large Language Models (LLMs) to become more aligned with human intentions in a scalable manner. Current strategies utilize large language models to annotate sets of model outputs to eventually align the model. However, there has not been much research done on the privacy preservation of the technique in sensitive environments such as in medicine, where private data cannot be leaked. To acheive this privacy preservation, a novel algorithm called *PriMa* is proposed for data obfuscation in the RLAIF pipeline, providing organizations with a way to preserve privacy and reap the benefits of RLAIF when utilizing externally hosted large language models. Specifically, this paper is one of the first to explore the privacy concerns in the context of membership inference attacks within practical implementations of RLAIF. The results of our experimental study show the improvement in defense against membership inference attacks, and a significant improvement in model alignment using our approach compared against the original RLAIF architecture.

CCS Concepts: • **Security and privacy → Domain-specific security and privacy architectures**; • **Computing methodologies → Natural language generation**.

Keywords: Reinforcement Learning from AI Feedback, AI Privacy, AI Alignment, Large Language Models, Masked Language Modeling

ACM Reference Format:
David Gao. 2023. Privacy Preserving RLAIF using Masking Algorithms. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In recent years, large language models (LLMs) have grown in popularity, as they have accelerated many tasks for many industries. However, training these LLMs takes a lot of resources, and not every company has the resources to take on this task, so the ownership of the largest language models lies within a couple organizations, the most notable being OpenAI [29]. Organizations may use these commercial LLMs, but must submit prompts through an API or an interface to interact with the LLMs which are hosted by the few large companies.

In order to keep their data in-house and customize models for their own needs, some organizations have started fine-tuning pre-trained models for their own needs, and organizations continue to try to leverage novel techniques to improve the usage of LLMs within their own institution. Specifically, many industries such as finance and healthcare, might have limited their use of commercial large language models due to privacy concerns, but may have an opportunity now to create better models for themselves with the private or proprietary data that are in their hands already.

Reinforcement Learning from AI Feedback (RLAIF) [2, 11] has been recently introduced as a new method in improving the capabilities of LLMs, overcoming the scaling limits which come from the use of human labelers provided in the original Reinforcement Learning with Human Feedback (RLHF) technique [19, 26]. One of the biggest discoveries is that these processes enable data efficient fine-tuning of pre-trained models which improves the alignment of a large language model to the intentions of humans [19], opening the door for users and organizations to create their own in-house models which are tailored to their own needs and data. This training pipeline has created models which vastly outperform supervised fine tuned models and base models across various methods of evaluation[11].

However, this approach has a bottleneck. The labeling process for the data which will be used to train the reward model cannot be easily scaled with human feedback, and RLAIF still requires the use of a large language model which evaluates internal data to implement the reinforcement learning pipeline. The cost of hosting a large language model on premise is high, so utilizing them for RLAIF would require usage of commercial APIs or cloud based solutions, releasing a portion of the data outside of the organization, undermining the purpose of training an in-house large language model in the first place. However, if we can find a way to leverage the

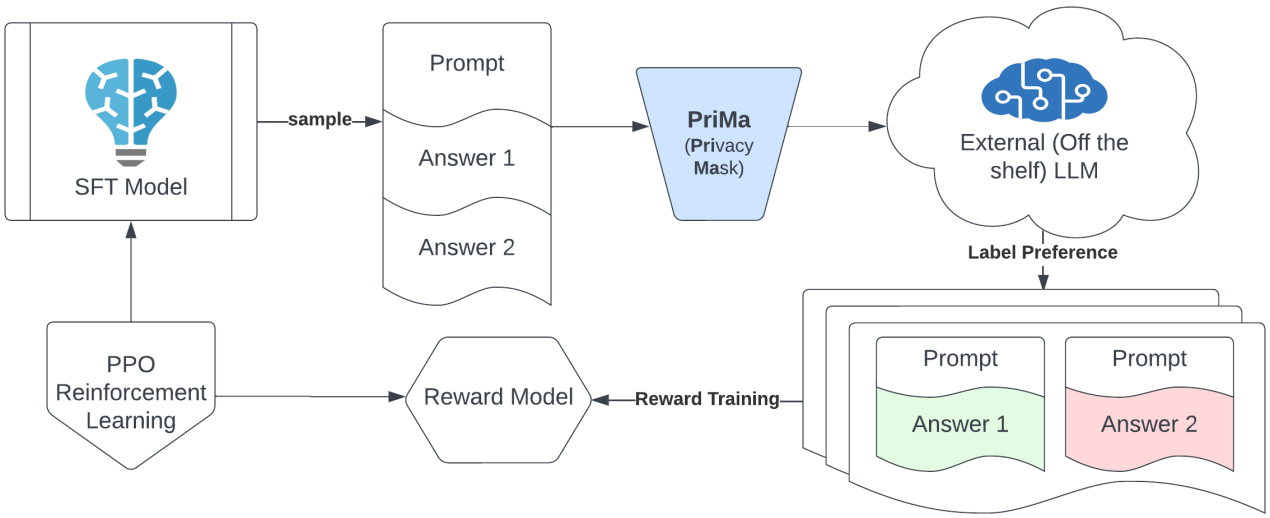

Figure 1. A diagram visualizing the RLAIF pipeline.

external LLMs in the RLAIF pipeline while preserving privacy of our reward model training data, organizations would be able to have better performing large language models for their own needs.

However, there hasn't been any extensive studies done in the area of privacy preservation for the RLAIF technique, with no known studies done up to the time of this writing. This remains new territory which needs much exploration.

In our solution, we propose the use of masked language models (MLMs)[5] to protect the privacy of an organization's data while maintaining the benefits of RLAIF when training their model using an externally hosted LLM. To protect against membership inference attacks [25] on the prompts we send to a commercial LLM in the RLAIF training pipeline, we use an algorithm to create synthetic sections of data to pass the external LLM, trying to optimize the preservation of privacy and the gains made from RLAIF.

The remainder of the paper is as follows. Section 2 goes over related work in the area of privacy preservation in natural language processing. Section 3 defines the RLAIF pipeline and the problem statement for the study. Section 4 presents our Masked Language Model algorithm for privacy preservation. Section 5 covers our experimental study. Section 6 presents the results and Section 7 concludes the paper.

2 Related Work

In this section, we go over related works that cover privacy preservation in the age of large language models. We also examine methods studied for anonymization and generation of datasets. Though there are no studies specific to privacy preservation for RLAIF, works in related fields provide necessary results which guide the direction of this study.

2.1 Large Language Model Privacy

During the meteoric rise of large language models, there have been many privacy concerns due to the vast amount of data these models hold and the power they can have due to widespread usage. One of the first papers which explores privacy concerns was written by Pan *et al.* which defines many privacy concerns and attack models that have become popular areas of research [20]. The biggest takeaway from this study is that during the usage of a commercial large language model, user information might be stolen from the prompts which are submitted to the model due to many possibly weaknesses. A lot of information can be revealed during inference time, even if an attacker does not have access to the model [20].

As supervised fine tuning became more popular for customizing LLMs for specific tasks, there was more research done on the information that could be extracted from such models, which might reveal information about the data it was trained on[6, 8, 9, 14, 15, 25, 27]. This was first generalized for broad machine learning models by Shokri *et al.*, an early work on membership inference attacks on models [25]. This study introduced the concept that many privacy risks exist when there is information to be gained from training data [25].

Specifically, Shokri *et al.* proposed the use of shadow models to imitate the target model as a way to successfully infer training set membership through the attack model[25]. Our experiments on privacy preservation are derived from this

concept and utilize it specifically for the data which is revealed as a part of the feedback process of the RLAIF pipeline.

2.2 Dataset Anonymization and Generation

There have been many studies on privacy preservation in data and the two main categories that I will focus on are the anonymization of data and generation of synthetic data, as the prominent approaches for privacy.

2.2.1 Anonymization. As the amount of data in the world increases, more concerns are brought up in the protection of Personal Identifiable Information (PII), and many works have explored strategies to remove this information from datasets. One of the earliest studies was done by Narayanan *et al.* in 2008 [17]. They prove that anonymization is not enough in many cases to protect the privacy of a user. By utilizing a few sources of information, algorithms can identify a user with small amounts of data [17]. Ohm built off this and continued to explore the limits of anonymization and evaluates future directions that should be taken in society [18].

Specifically, applications to the medical field are discussed, which remains a point of focus in the present, among other fields. Many studies go utilize modern machine learning and deep learning models to anonymize data, and begin to automate the process, making anonymization a more scalable technique for privacy preservation of a dataset [4, 16, 21]. However, as we can infer from the earlier studies by Narayanan *et al.* and Ohm, this anonymization might not be enough, especially with new techniques and models that attackers can use to gather private information from this anonymized data.

2.2.2 Generation. Generation of synthetic data which imitates the real dataset is an approach that has gained traction in recent years. There have been many different approaches, but we will focus on the Masked Language Modeling (MLM) concepts proposed by Devlin *et al.* in the context of text generation and augmentation. MLMs have been used to create synthetic data successfully in various studies, and have proved to be an effective way to preserving privacy of a dataset while maintaining the same characteristics of the dataset [28, 30]. Data from synthetic data has even been utilized by Kweon *et al.* recently to train an open source large language model, proving that synthetically generated data can be adequate for LLM training [10].

3 Problem Statement

We establish examples and definitions in this section that will serve as a reference for the rest of the paper, and we define our problem statement from it.

Definition 3.1. Pre-trained Large Language Model: A pre-trained large language model is a model which already has learned weights from a large corpus of data. These models can usually be accessed by API from a company that hosts it. As an alternative, there are open sourced models which can be deployed locally, if enough resources are in place. Typically, these models are general purpose and can serve a wide variety of needs.

Definition 3.2. Supervised Fine-Tuning (SFT): We define supervised fine-tuning as the process where you can use a smaller scale data set to enhance the capabilities of a pre-trained LLM and have it better serve specific needs. This process isn't as computationally expensive as training an LLM from scratch, and requires less data [5]. These models are typically trained and utilized for downstream tasks such as sentiment analysis, question answering, and classification.

Definition 3.3. One-shot RLAIF: One-shot RLAIF is the process proposed by Lee *et al.* [11] which uses an off the shelf LLM to rank preference for outputs given by a model to a given prompt. A prompt, and a collection of responses generated by the SFT model (two in this study) is passed into an off the shelf LLM along with evaluation guidelines. The LLM outputs a preference, providing a data labeling pipeline to train the reward model for reinforcement learning. Each data point passed into the external LLM has at least three designated sections:

1. The Prompt: A text sequence that is to be inputted into the original model.
2. Answer 1: A sample answer to the prompt (typically generated by the original model).
3. Answer 2: A second answer to the prompt.

The external LLM is then instructed to select the answer which is "better", typically in terms of helpfulness or some other arbitrary measurement.

Example 3.4. Headlands Hospital wants to create an in-house large language model to assist their doctors in the process of diagnosing symptoms of a patient. They want to train a model by fine-tuning a model using the vast amounts of data they have within their systems, but realize that the performance isn't up to their standards, so they propose to use Reinforcement Learning from Artificial Intelligence Feedback to improve their model. Specifically, they hope to not only increase accuracy of responses, but to improve the helpfulness of responses in regard to user intention. They would love to use GPT 4 to help train their model, since it is one of the largest LLMs on the market, and boasts impressive performance across multiple modalities. However, they don't want to accidentally leak any private data and risk violation of any HIPAA policies.

Example 3.5. Fantastic Finance is a financial institution who wants to host a large language model to help their customer support employees fill out repetitive tasks with the help of the LLM. They have many documents within their organization that they want to use to fine-tune the model, but they want it to hallucinate less, since inaccurate

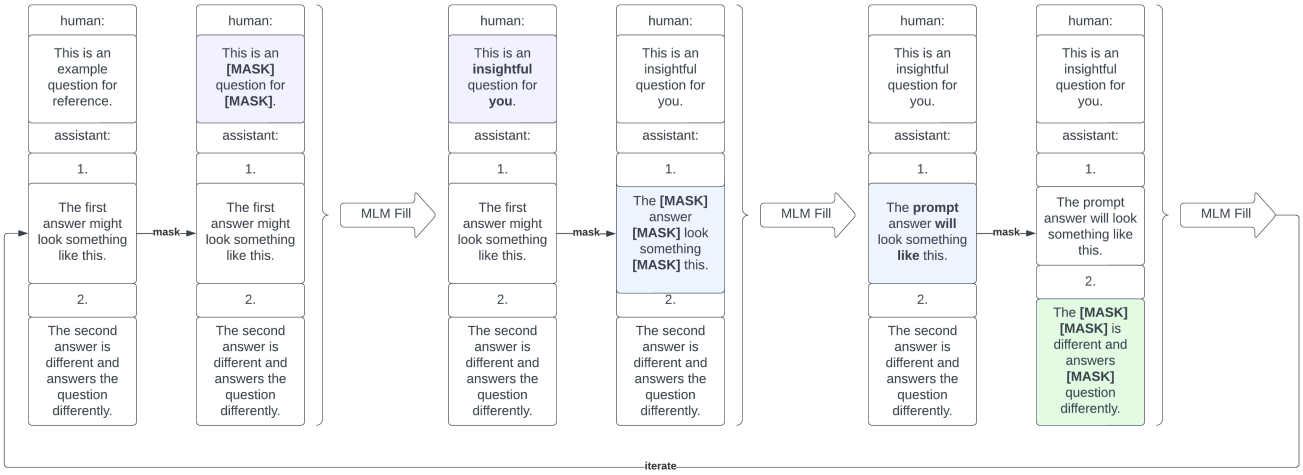

Figure 2. A simplified step by step of PriMa.

information could cost their firm millions. They decide to use RLAIF to improve their model, but don't want to expose proprietary and customer information to possible attackers.

3.1 Problem Definition

With the scenarios given in mind, we can define what we want from a privacy preserving RLAIF pipeline. This can be summarized from two points.

1. **Privacy Preservation:** We want to ensure that the information that exists outside the organization cannot be linked back to private data which is learned by the SFT model. We propose the the biggest risk in this setting is **membership inference**. To measure this, we will use a modified version of the shadow attacks proposed by Shokri *et al.* [25] which we will define in the experimental study section of this paper.
2. **Training Effectiveness:** We also want to measure the effectiveness of the RLAIF training on the performance of the model being trained.
   - To do this, we will hold out a test dataset, and compare the capabilities of the SFT model, the vanilla RLAIF model, and the privacy preserving RLAIF model using the various capabilities of the ROUGE evaluation first proposed by Lin [12].
   - We also would like to measure the alignment of the model. This will be done with pairwise comparisons of answers, using another model to evaluate the preferred response. This is similar to the work done by Lee *et al.* in their comparisons of RLAIF models to other models.

These two measures will have expected tradeoffs, which we will examine in our experimental study, and can be a future research direction.

Algorithm 1 PriMa Algorithm

Require: $input, iterations, proportion$
for $i = 1$ **to** $iterations$ **do**
  $prompt, answer1, answer2 \leftarrow \text{Split}(input)$
  $sections \leftarrow \{prompt, answer1, answer2\}$
  for all $section$ in $sections$ **do**
    for all $word$ in $prompt$ **do**
      if Random() < $proportion$ **then**
        Replace $word$ with **[MASK]**
      end if
    end for
    Join($prompt, answer1, answer2$)
    Fill [MASK] Tokens
  end for
end for
return Join($prompt, answer1, answer2$)

4 PriMa Algorithm

The Privacy Mask (PriMa) algorithm augments a data point specifically for RLAIF. Each token has an independent **M%** of being replaced within each section of the data (consisting of a prompt and two generated responses), and we iterate through the data **N** times. The goal of this is to provide M and N as adjustable parameters to balance the trade off between privacy preservation (token replacement) and content preservation for the RLAIF labeler (coherence). The general pseudocode is provided in Algorithm 1.

4.1 Probabilistic Mask and Replacement

To preserve the overall correctness of the output, we extract the individual sections from our input data, keeping the overall structure to ensure accurate labeling. Within the first section (the prompt), we iterate through each word, and each

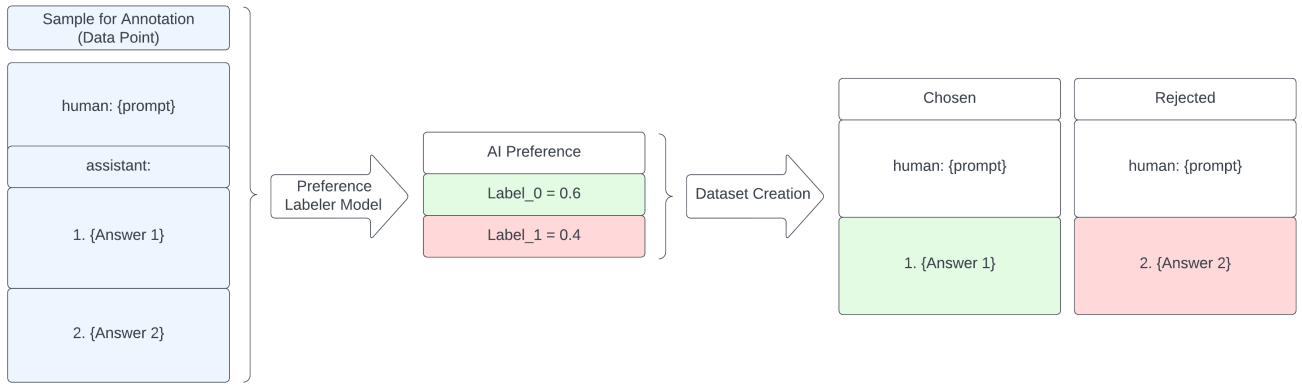

Figure 3. A diagram for the AI preference labeling process.

has an M% chance of being replaced by a **[MASK]** token. Then, all three parts are passed together into a TinyBERT [7] model to obtain the most likely token for all masked tokens. The masked tokens are replaced by the highest probability token determined by the TinyBERT model, and the sections are separated again. This process is then repeated for the other sections (Answer 1 and Answer 2).

By separating the text into sections and masking them separately, we ensure that the overall text remains coherent. This enables the external LLM to effectively label our data, and it also guarantees that we have high quality text for training our reward model.

Also, since the algorithm has a random chance of selecting each word, we maintain a higher level of privacy, since it makes it more difficult for attackers to reverse engineer the algorithm and uncover the replaced sections of text.

4.2 Iteration

Using the probabilistic replacement, we can iterate through the entire input N times, increasing the amount of replaced tokens each time. Any amount of iterations should maintain stability since only a certain amount of the total data is masked at a time. However, for each iteration, the privacy for the input data should increase, and the output data will have diminishing resemblance.

In theory, there is an asymptote for the effectiveness of iteration, since the masked language model will maintain the general structure of each section, but will eventually replace all tokens in the data such that any mask will result in the token being replaced by the original word.

5 Experimental Study

The experimental study comprises the majority of this study, since multiple RLAIF trained models are needed to evaluate the effectiveness of our PriMa Algorithm.

5.1 Dataset Splitting

The MedQuAD dataset [3] published 16.4K publically available pairs of question and answer pairs to medical questions which we use throughout this study. We split this up into the following sections:

1. 5.74K questions and answer pairs to fine-tune our target SFT
2. 5.74K questions and answer pairs to fine-tune our shadow model.
3. 820 out-of-sample questions for the Vanilla RLAIF pipeline.
4. 820 out-of-sample questions for the shadow classification task.
5. 1.64K questions designated for reinforcement learning.
6. 1.64K questions and answers for testing.

5.2 Mock Preference Labeler Model

To keep the experimental study within a controlled environment and preserve resources, we did not use a commercial LLM to label the preferences coming out of the SFT. We fine tuned a DistilBERT model [23] for text sequence classification based on the hh-rlhf dataset [1] curated by Anthropic. This dataset emphasizes the "helpfulness" of a model's responses [1]. We took the hh-rlhf dataset and turned it into a binary classification task. Given the prompt and the two answers, the target output is "0" if the first answer is preferred and "1" if the second answer is preferred.

We utilize the model as a substitution for the "External LLM" in Figure 1 to create the training set for the Vanilla Reward Model and the Privacy Preserving Reward Model. In our evaluations, we also reuse the model as a grader to determine win rates of models, head to head, when compared on test prompts. This will be explored in our result section under Alignment Increase.

The model is available at https://huggingface.co/davidgaofc/hh-labeler.

5.3 Target Supervised Fine-Tuned Model (Target SFT)

As a base for RLAIF, we fine-tuned a t5-small model [22] on a portion of the MedQuAD dataset [3] which provides medical question and answer pairs from various sources. We leave out other portions of the MedQuAD dataset for further usage throughout the study.

This sequence to sequence serves as a baseline for the RLAIF training process and also acts as the target model in our privacy evaluations on membership inference attacks and is able to provide general answers to simple medical questions.

The model is available at https://huggingface.co/davidgaofc/SFT_Med_t.

5.4 Vanilla RLAIF Pipeline

In order to have a baseline comparison for our specific task, we go through the entire process of training a model using RLAIF, originally proposed by Lee *et al.* [11] in their groundbreaking paper.

5.4.1 Vanilla Reward Model (Vanilla RM). We extract pairs of responses to questions which come from another portion of the MedQuAD dataset to train our reward model in the typical RLAIF pipeline.

These are mixed in with an equal number of randomly sampled in-sample prompts (prompts that were used in training). This provides a balanced dataset for our attack model described later and represents a worst-case scenario for real life. In practice, the percentage of prompts that reveal private information is generally unknown due to the somewhat stochastic nature of generative artificial intelligence.

These pairs of responses come directly from the Target SFT specified previously with varied temperature to ensure the model gives two different responses. Then, the questions and responses are formatted together and passed through the mock preference labeler model mentioned previously.

After they are labelled, they are split into another dataset where the columns are "chosen" and "rejected", where the "chosen" column contains the prompt and the response selected by the labeler and the "rejected" column contains the prompt and the response not selected.

Using this dataset, we train fine tune a base DistilRoBERTa model [13, 23] for a single label text classification task so that it outputs a single score which we can use for our reinforcement learning signal later.

This reward model is available at https://huggingface.co/davidgaofc/RM_base.

5.4.2 Shadow Supervised Fine-Tuned Model (Shadow SFT).

5.4.3 Vanilla RLAIF Model. Using the Vanilla RM from the previous section, we utilize the Proximal Policy Optimization (PPO) Algorithm proposed by Schulman *et al.* [24] to align our Target SFT. This is done using prompts from another subsection of the MedQuAD dataset designated for the reinforcement learning part of our pipeline.

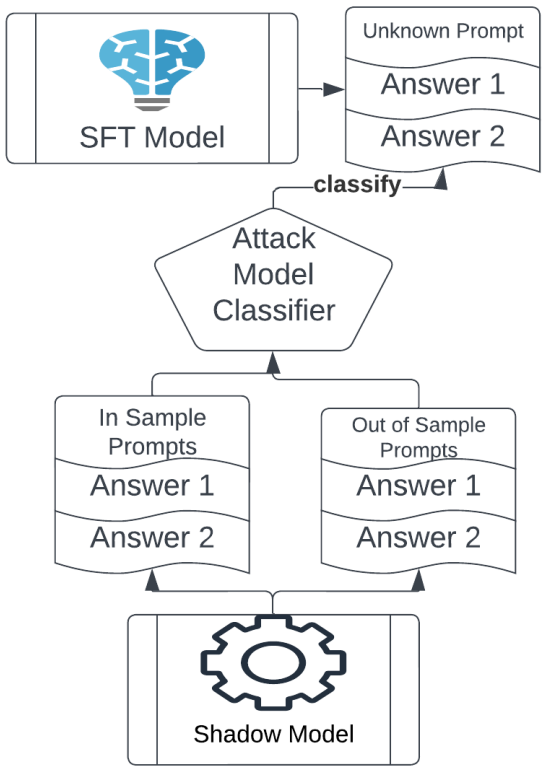

Figure 4. A diagram showing the Shadow attack for membership inference.

This is the final product of the Vanilla RLAIF pipeline which we will use to compare against the model that comes out of our privacy preserving RLAIF pipeline.

This model can be found at https://huggingface.co/davidgaofc/PPO_base.

5.5 Privacy Preserving RLAIF Pipeline

We begin a separate process of privacy preserving RLAIF as shown in Figure 1. This mirrors the Vanilla RLAIF pipeline with the addition of our novel PriMa algorithm.

5.5.1 Privacy Preserving Reward Model (Privacy RM). Using the same pairs of responses generated for the Vanilla RM specified previously (both in-sample and out-of-sample), we pass the dataset through our PriMa algorithm, masking with a probability of 30% and iterating once through the data.

This gives us a privacy preserving dataset, which is passed to the mock preference labeler model. After this dataset is labeled, we take it and split it into a dataset fit for training the reward model, splitting datapoints into "chosen" and "rejected" using the same process as the Vanilla RM.

This dataset now contains "masked" responses which no longer directly resemble the original outputs of the Target SFT, but remains the same size as the original dataset used to train the Vanilla RM.

We utilize this dataset to fine-tune a base DistilRoBERTa model (same base model as the Vanilla RM) for single label text classification, using the exact same training parameters as the Vanilla RM to maintain comparability.

This model is available at https://huggingface.co/davidgaofc/RM_prima.

5.5.2 Privacy Preserving RLAIF Model. Using the Privacy Preserving Reward Model as a reinforcement learning signal, we utilize the Proximal Policy Optimization (PPO) algorithm on our Target SFT once again, using the same prompts designated for the reinforcement learning process from the MedQuAD dataset. This time, we utilize the Privacy RM as the reinforcement learning signal, and we utilize the same training parameters as we did for the Vanilla RLAIF model.

This is the final product of the privacy preserving RLAIF pipeline, and is used for comparison against both the base SFT and the Vanilla RLAIF model in our results section.

This model is available at https://huggingface.co/davidgaofc/PPO_prima.

5.6 Shadow Membership Inference Attack

In order to measure the privacy preservation of our PriMa algorithm, we implement the concepts of Shokri *et al.* and frame it in reference to the data labeling step in the RLAIF pipeline [25]. As of the writing of this paper, we have not seen any other studies specific to this process, so we don't have a baseline for comparison.

The goal of the shadow attack is to train a separate model on similar distributions of data. Then, the attacker can create a dataset on the model's responses to in-sample and out-of-sample prompts. Using this labeled dataset, they can try to infer membership on a target model. This process is visualized in Figure 4. In practice there are varying degrees of success with this strategy since the attack depends completely on the data chosen, but the primary measure is the precision of the attack [25].

The first step in implementing this attack is to train a model which ideally has the same architecture as the target model, and is trained on a disjoint dataset.

We fine-tune a t5-small model on a held out portion of the MedQuAD dataset that is disjoint to the dataset that our target model (the Target SFT) is trained on.

We also test it on the same test dataset as the target model to guarantee similar overall performance. In practice, attackers can obtain these models or train them, depending on the situation, but we pessimistically assume that an attacker can replicate similar results due to the vast amounts of data which is publicly available.

	Base Data	PriMa Data
Attack Precision	50.19	49.37

Table 1. Membership Inference Attack Results

This shadow model is available at https://huggingface.co/davidgaofc/SFT_shadow.

Using this model, we create a training dataset for our attack model. We prompt the shadow model for two responses, mirroring the way we prompted the target SFT in the beginning of Section 5.3.1. The out-of-sample prompts are taken from another disjoint segment of the MedQuAD dataset that we held out for this purpose.

This creates a labeled dataset for classification, where the goal is to classify whether a given datapoint comes from within the training set or is out-of-sample. In our next section, we examine the architecture for this membership inference attack model.

5.6.1 Attack Classifier Model. We fine-tune a base DistilBERT model for text classification to act as our attack model in this study.

Using the labeled dataset mentioned in the previous section, we split it into training (80%) and validation (20%) datasets to prevent overfitting.

This model is available at https://huggingface.co/davidgaofc/ShadowAttackF.

After fitting this model on the text classification task, we test it on the data used for the Vanilla RLAIF pipeline and the obfuscated data from the Privacy Preserving RLAIF pipeline. We will analyze these results in the next section.

6 Results

In this section, we go over the results of our experimental study.

6.1 Membership Inference Privacy

As mentioned previously, the metric used for calculating the effectiveness of a membership inference attack is the precision. This is able to measure the percentage of in-sample data points that the attack model is able to identify correctly.

As you can see in Table 1, the original data (base data) used for the Vanilla RLAIF pipeline already has a very low attack precision. We attribute this to the complexity of the classification problem and the limitations posed by the amount of data we used to train the attack model.

However, we do see a decrease in the precision after the data is passed through the PriMa algorithm after a single iteration with 30% likelihood of token masking, indicating that our algorithm hides in-sample data points to a degree.

In practice this decrease can be significant. In sensitive environments like medicine and finance, membership inference of a single datapoint could result in many issues for

	SFT	Vanilla RLAIF	Privacy RLAIF
ROUGE-1	26.79	26.01	22.77
ROUGE-2	11.95	12.11	12.49
ROUGE-L	22.09	21.48	19.32

Table 2. ROUGE scores for each model on test dataset.

organizations. The precision decrease below 50% could deter attackers enough to provide a higher level of privacy for these organizations.

6.2 ROUGE Stability

Through all of this, we want to ensure that the language models still provide the best answers to the questions when utilized by the organizations. Our first evaluation are the ROUGE metrics [12] for comparing responses to correct reference responses in our test dataset.

As you can see in Table 2, we maintain stability across most of the ROUGE metrics for all three of the models.

We do see a significant decrease in the ROUGE-1 score for the privacy preserving RLAIF, but this is expected. By masking and replacing tokens in the reward model, the final model will inherently move away from certain words in its vocabulary. This is an expected side effect of the privacy preservation, but is not concerning, since the absence of single words does not necessarily equate to worse responses.

The ROUGE-2 score provides a balanced perspective for evaluation. We see that both RLAIF models are able to string together two-word phrases more similar to the reference text than the SFT. This negates our concern from the ROUGE-1 scores, and indicates that our RLAIF models can still answer questions with similar accuracy as the SFT model.

The ROUGE-L score indicates that both RLAIF models generate responses that are less structurally similar to the reference text than the SFT model, but this can be attributed to the results of the PPO algorithm, and do not necessarily indicate worse responses.

6.3 Alignment Increase

Since the main benefit of RLAIF is the alignment it provides to the model, we must measure the result of our novel algorithm on this measurement.

To do this, we reuse the mock preference labeler model from Section 5.2 to annotate head to head responses to the test dataset. Using this, we calculate the win rate of model responses against one another for a given prompt. If a model's response "wins", it means the answer is more aligned, and in this case, more "helpful" than the other answer.

As we can see in Figure 5, both the Vanilla RLAIF model wins the majority of the time against the SFT, replicating the results of Lee *et al.* to a lesser degree [11].

In our evaluations, we see that the model that has gone through the privacy preserving RLAIF pipeline (using our PriMa algorithm), has an even higher win rate against the SFT than the Vanilla RLAIF algorithm. When compared head to head, the AI preference labeler prefers our model's responses the majority of the time compared to the Vanilla RLAIF algorithm as well.

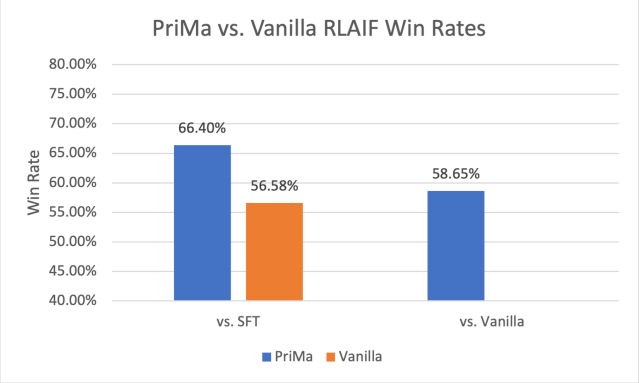

Figure 5. Results from head to head annotations by the AI Preference Labeler.

We theorize that the preprocessing step done by our PriMa algorithm generates a less noisy sample for the preference labeler to annotate, resulting in higher quality data for the reward model to be trained on. We leave this for future studies to expand on, as it signifies exciting possibilities for RLAIF alignment improvement.

7 Conclusion

In this word, a novel approach for preprocessing data in the Reinforcement Learning from AI Feedback (RLAIF) is evaluated. The PriMa algorithm utilizes a unique segmentation and iteration process to augment data points using Masked Language Modeling (MLM) before they are labeled by an external LLM in the RLAIF training pipeline. The proposed approach improves the privacy preservation of the released prompts, decreasing the precision of shadow membership inference attacks. It also improves the alignment ability of the final model while maintaining accuracy after the entire RLAIF training pipeline.

Acknowledgments

To Dr. Dan Lin, for her guidance and mentorship throughout the process of this study. Her wisdom and motivation kept this project on track and ensured the completion of the paper. To Dr. Jesse Spencer-Smith, for inspiring me to explore the capabilities of transformer models and teaching me so much in his course. To my family and friends for supporting me through my first end-to-end research experience, and helping me push through, even in moments of self-doubt.

References

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[3] Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics* 20, 1 (2019), 1–23.

[4] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24, 3 (2017), 596–606.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305* (2021).

[7] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351* (2019).

[8] Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. 2023. User Inference Attacks on Large Language Models. *arXiv preprint arXiv:2310.09266* (2023).

[9] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881* (2023).

[10] Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. 2023. Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes. *arXiv preprint arXiv:2309.00237* (2023).

[11] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267* (2023).

[12] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[14] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *arXiv preprint arXiv:2302.00539* (2023).

[15] Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv preprint arXiv:2203.03929* (2022).

[16] Suntherasvaran Murthy, Asmidar Abu Bakar, Fiza Abdul Rahim, and Ramona Ramli. 2019. A comparative study of data anonymization techniques. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 306–309.

[17] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.

[18] Paul Ohm. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.* 57 (2009), 1701.

[19] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155* 13 (2022).

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

[21] Constantinos Patsakis and Nikolaos Lykousas. 2023. Man vs the machine: The Struggle for Effective Text Anonymisation in the Age of Large Language Models. *arXiv preprint arXiv:2303.12429* (2023).

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[25] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[26] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[27] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 268–282.

[28] Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2022. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348* (2022).

[29] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[30] Nina Zhou, Qiucheng Wu, Zewen Wu, Simeone Marino, and Ivo D Dinov. 2022. DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes. *Journal of Medical Systems* 46, 12 (2022), 96.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009