2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA) | 979-8-3503-8674-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/TPS-ISA62245.2024.00033

Preserving Privacy During Reinforcement Learning With AI Feedback

David Gao
Department of Computer Science
Vanderbilt University
Email: david.gao@vanderbilt.edu

Ian Miller
Department of Computer Science
Vanderbilt University
Email: ian.miller@vanderbilt.edu

Ali Allami
Department of Computer Science
Vanderbilt University
Email: ali.allami@vanderbilt.edu

Dan Lin
Department of Computer Science
Vanderbilt University
Email: dan.lin@vanderbilt.edu

***Abstract*—Leveraging the scalable efficacy of reinforcement learning from AI feedback (RLAIF), large language models (LLMs) can be refined toward human intent alignment. While current paradigms employ LLMs to annotate and refine model outputs, the area of privacy preservation within RLAIF, particularly in sensitive domains like healthcare where data privacy is indispensable, remains inadequately explored. Addressing this gap, we propose *PriMa*, a novel data obfuscation algorithm integrated into the RLAIF pipeline, empowering organizations to harvest the benefits of RLAIF while protecting privacy concerns. Notably, this work pioneers the investigation of privacy vulnerabilities, specifically membership inference attacks, within practical RLAIF implementations. Our empirical evaluations demonstrate PriMa's effectiveness in preventing membership inference attacks while significantly enhancing model alignment compared to the baseline RLAIF architecture.**

I. Introduction

The thriving prominence of large language models (LLMs) has significantly enhanced the efficiency of diverse tasks across numerous industries. However, the substantial resource demands associated with LLM training have concentrated ownership of the most potent models within a select few entities, with OpenAI serving as a premier example [1]. This centralization necessitates external organizations to interact with LLMs hosted on large-scale infrastructure via APIs or interfaces, forfeiting direct control over data processing.

Recognizing the value of in-house data privacy and customization, some organizations have opted to fine-tune pretrained models for their specific needs. Furthermore, continued efforts in novel techniques aim to optimize LLM utilization within individual institutions. Notably, privacy concerns within domains such as finance and healthcare may have previously restricted reliance on commercial LLMs. However, the availability of proprietary data presents these sectors with an invaluable opportunity to develop and deploy superior models tailored to their unique requirements.

Emerging within the domain of LLM advancements, Reinforcement Learning from AI Feedback (RLAIF) [2], [3] addresses the scalability limitations inherent in the predecessor methodology, Reinforcement Learning with Human Feedback (RLHF) [4], [5]. By leveraging AI-generated feedback instead of human labelers, RLAIF demonstrably facilitates data-efficient fine-tuning of pretrained models [4]. This refinement enhances the alignment of LLMs with human intent, empowering users and organizations to cultivate in-house models tailored to their specific needs and data. Extensive evaluations [3] reveal that RLAIF-trained models significantly outperform their supervised fine-tuned and base model counterparts across diverse assessment metrics.

Despite RLAIF's efficacy in achieving data-efficient fine-tuning and human intent alignment, it encounters a crucial bottleneck in the labeling process for its reward model. Scaling this process via human feedback remains challenging, and the technique still necessitates a large language model (LLM) for internal data evaluation within the reinforcement learning pipeline. Hosting such an LLM on-premise incurs significant costs, and relying on commercial APIs or cloud-based solutions for RLAIF implicates data exfiltration, diminishing the very purpose of in-house LLM development. Therefore, organizations desire a mechanism to harness the effectiveness of external LLMs within the RLAIF pipeline while simultaneously protecting the privacy of their reward model training data. Achieving this accomplishment would pave the way for organizations to cultivate superior LLMs tailored to their specific needs and data, unlocking the full potential of the RLAIF paradigm.

While RLAIF has sparked significant academic discourse regarding its efficacy in data-efficient fine-tuning and human intent alignment, a noticeable gap exists in the area of privacy-preserving methodologies. No prior studies have investigated this critical aspect, highlighting an urgent need for comprehensive research initiatives to equip organizations with the tools to leverage RLAIF's advantages while ensuring the confidentiality of sensitive data within the reward model training process.

To circumvent privacy concerns associated with the RLAIF training pipeline, we propose leveraging masked language

979-8-3503-8674-5/24/$31.00 ©2024 IEEE
DOI 10.1109/TPS-ISA62245.2024.00033

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

models (MLMs) [6] for data obfuscation while preserving the efficacy of model fine-tuning. Specifically, we address the vulnerability of prompts transmitted to external LLMs (membership inference attacks, as identified by [7]) by incorporating an algorithm that generates synthetic data segments. This technique optimizes a privacy-utility trade-off, aiming to maximize privacy protection while minimizing potential performance degradation in the RLAIF process.

In this paper, we propose a novel algorithm using a masking technique to preserve the privacy of sensitive data in the RLAIF model against the membership interference attack. Specifically, we consider the RLAIF model where organizations aim to create an in-house large language model to assist their staff in the process of task analysis. We summarize two realistic examples that align with this scenario as follows:

Headlands Hospital seeks to augment its diagnostic capabilities by developing an in-house large language model (LLM). Leveraging their extensive internal patient data, they propose fine-tuning a pre-trained model to enhance its performance beyond current benchmarks. The proposed approach entails Reinforcement Learning from AI Feedback (RLAIF) to optimize not only the accuracy of the LLM's responses but also their alignment with user intent - a crucial factor in patient care. While the prospect of harnessing the exceptional capabilities of GPT-4, a leading LLM in terms of size and multi-modal prowess, is enticing, privacy concerns regarding potential HIPAA violations necessitate a more nuanced approach. Therefore, exploring innovative techniques to safeguard sensitive patient data while enabling effective RLAIF-driven LLM training becomes paramount for Headlands Hospital's endeavor.

Fantastic Finance aims to leverage the automation potential of large language models (LLMs) to alleviate the burden of repetitive tasks facing its customer support team. The company possesses a rich internal document repository ideal for model fine-tuning, yet mitigating the generation of "hallucinations" (factually inaccurate outputs) is crucial, as financial missteps could incur significant losses. Recognizing the efficacy of Reinforcement Learning from AI Feedback (RLAIF) in tailoring LLM behavior, Fantastic Finance seeks to implement this technique while simultaneously addressing critical privacy concerns. Ensuring the confidentiality of proprietary and customer data during the RLAIF process becomes a cornerstone of their endeavors, necessitating the exploration of privacy-preserving techniques to protect sensitive information from potential attackers.

This work examines the intricate world of Reinforcement Learning from AI Feedback (RLAIF), proposing and examining a novel data preprocessing approach named *PriMa*. *PriMa* depends on a unique blend of data segmentation, iterative masking, and Masked Language Modeling (MLM) to accurately augment data points before their crucial labeling stage by an external LLM within the RLAIF pipeline.

In short, our main contributions can be outlined as follows:

1) **Enhanced Privacy Preservation:** We introduce the *PriMa* algorithm, which utilizes a unique combination of data segmentation, iterative masking, and Masked Language Modeling (MLM) to obfuscate sensitive information in training prompts. We demonstrate that *PriMa* somewhat reduces the precision of membership inference attacks, thereby further protecting the privacy of training data. Our result holds significant practical implications for organizations deploying RLAIF models in sensitive domains by mitigating the risk of unauthorized data inference.
2) **Improved Model Alignment:** We propose that *PriMa* augmented data improves the alignment of RLAIF models with desired characteristics by providing cleaner training examples for the external LLM labeler. Our comprehensive evaluation reveals that models trained with *PriMa* augmented data exhibit a notable improvement in their ability to generate responses aligned with specific criteria compared to models trained with vanilla RLAIF. This finding signifies the potential of *PriMa* to promote the development of RLAIF models that not only deliver accurate outputs but also align with specific application requirements.

The rest of the paper is organized as follows: Section II surveys relevant prior work in the domain of privacy-preserving natural language processing (NLP) techniques. Section III formulates the problem statement guiding this study. Section IV presents our proposed algorithm. Section V elaborates our experimental methodology. Section VI reports the experimental results. Finally, Section VII concludes the paper.

II. Related Work

This section focuses on a detailed survey of relevant prior works addressing the vital challenge of privacy preservation in the burgeoning era of large language models (LLMs). We explore existing research on diverse anonymization techniques and dataset generation methodologies. While studies directly focused on RLAIF-specific privacy preservation remain elusive, valuable insights and transferable results gleaned from the broader NLP and machine learning domains illuminate the path forward for this investigation.

A. Large Language Model Privacy

The growing importance of large language models (LLMs) has ignited a plethora of privacy concerns, stemming from their voracious data appetites and potent capabilities facilitated by ubiquitous deployments. Pioneering work by Pan et al. [8] outlines a spectrum of privacy threats and malicious attack models, inspiring subsequent research endeavors in these domains. A critical takeaway from their study is the inherent vulnerability of user information embedded within prompts submitted to commercially hosted LLMs. Exploiting potential weaknesses in these prompts, even without direct model access, attackers can glean significant amounts of sensitive data during inference time [8].

The surge in popularity of supervised fine-tuning for task-specific customization of LLMs prompted concurrent progress

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

in research concerning information leakage potential. Several studies [7], [9]–[14] shed light on the extractable information from such models, potentially revealing details about the underlying training data. Shokri et al. [7] laid the groundwork with their seminal work on membership inference attacks (MIAs) in machine learning models, generalizing the concept beyond LLMs. Their groundbreaking study established the inherent privacy risks associated with information leakage from training data, paving the way for subsequent research into various MIA techniques and mitigation strategies [7].

Shokri et al.'s [7] groundbreaking work on membership inference attacks (MIAs) laid the foundation for our privacy-preserving approach within the RLAIF context. Their seminal paper introduced the concept of ”shadow models,” alternate models trained to mimic the target model's output, enabling successful membership inference through attack simulations [7]. Inspired by this innovative technique, our experiments leverage the shadow model concept to specifically analyze the information leakage potential within the RLAIF pipeline. Duan et al. [15] found that MIAs are generally not very effective against LLMs, which is mostly consistent with our findings, however we do still see some reduction in attack efficacy. If it should happen at some point in the future that a more viable MIA against LLMs comes out, the method we propose here will likely be effective against that too.

B. Dataset Anonymization and Generation

The area of data privacy preservation has been a productive ground for rigorous research, yielding a range of promising approaches. This study examines two prominent paradigms within this domain: data anonymization and synthetic data generation, each offering distinct advantages and challenges in safeguarding sensitive information.

1) Anonymization: Focusing on applications in the healthcare domain, a contemporary area teeming with privacy concerns, numerous investigations leverage modern machine learning and deep learning architectures for data anonymization. They strive to automate the process and enhance its scalability for robust dataset privacy preservation [16]–[18]. However, as exemplified by the pioneering work of Narayanan et al. and Ohm, such anonymization alone may not suffice, particularly in the face of evolving adversarial techniques and model capabilities that attackers can exploit to glean private information from anonymized data.

2) Generation: In the domain of privacy preservation, a thriving paradigm has emerged: the synthesis of data that faithfully replicates the intricacies of real-world datasets. This strategy offers a compelling solution to the inherent tension between data utilization and individual privacy. Notably, a plethora of diverse approaches have arisen, each competing for efficacy in this domain. Among these, Masked Language Modeling (MLM) stands out as a particularly promising avenue, drawing inspiration from its groundbreaking application in text generation and augmentation as pioneered by Devlin et al. [6].

The unquestionably adept capabilities of MLMs in synthetic data generation have been showcased in numerous studies. These endeavors have successfully safeguarded sensitive data and preserved the essential characteristics required for downstream analytical endeavors [19], [20]. For instance, Kweon et al.'s groundbreaking work in synthesizing data for training an open-source LLM provides evidence of the viability and efficacy of this approach in the context of LLM development [21]. This success story formalizes the foundation for further exploration and refinement of synthetic data generation techniques, promising a future where data-driven insights can be gleaned without compromising the fundamental right to privacy.

III. Preliminaries

To lay the groundwork for the subsequent discourse, this section defines key terms that will serve as essential reference points throughout the remainder of this paper.

A. One-shot RLAIF

A one-shot RLAIF processes a tuple $(P, A1, A2, E, R)$ where P is the prompt, represented as a text sequence x_p. $A1$ and $A2$ are two potential responses generated by the Supervised Fine-Tuning (SFT) model, represented as text sequences x_1 and x_2. E is the evaluation guidelines, encoded as a set of instructions I for the external LLM. R is the external LLM, represented as a function $R(x_p, x_1, x_2, I)$ that outputs a preference score in the range $[0, 1]$, indicating the relative ”betterness” of $A1$ compared to $A2$ based on the prompt and evaluation guidelines. Figure 1 illustrates the AI preference labeling process in general.

The one-shot RLAIF process is as follows:

1) **Gather data points:** Construct a dataset $D = (P_i, A1_i, A2_i, E)$ for $i = 1, ..., N$.
2) **Query external LLM:** For each data point in D, obtain $R(P_i, A1_i, A2_i, E)$.
3) **Train reward model:** Use the preference scores from R to train a reward function $r(P, A)$ that approximates the external LLM's evaluation of a response A to a prompt P.
4) **Reinforcement learning:** Employ the learned reward function $r(P, A)$ to guide reinforcement learning algorithms in fine-tuning the SFT model to generate responses that align with the desired preferences.

B. Problem Definition

Drawing upon the presented scenarios, we now formally define the desired attributes of a privacy-preserving RLAIF pipeline. This requirement can be clarified into two fundamental points.

Privacy Preservation: Our primary objective is to preserve the privacy of sensitive information learned by the SFT model. This entails preventing the inference of whether specific data points from outside the organization (denoted as $I_{external}$) belong to the model's private training dataset ($D_{private}$). We hypothesize that membership inference attacks pose the

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

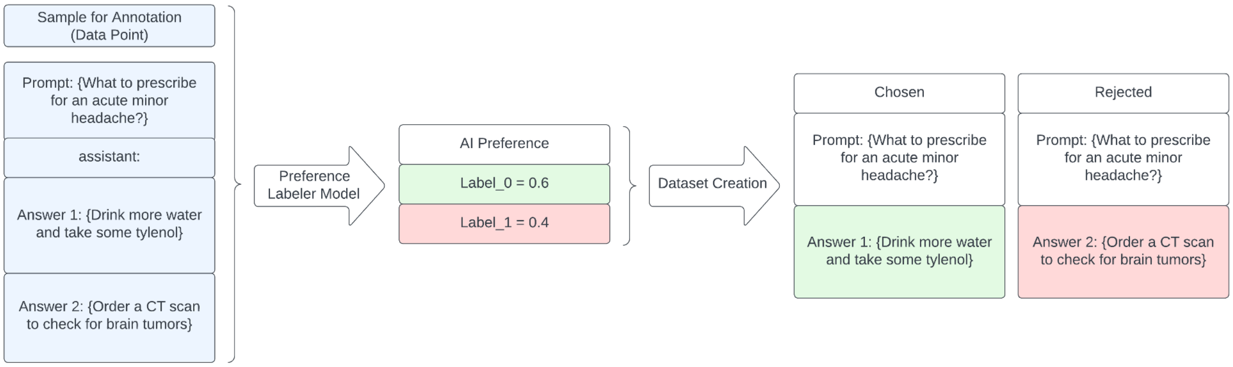
Sample for Annotation (Data Point)
Prompt: {What to prescribe for an acute minor headache?}
assistant:
Answer 1: {Drink more water and take some tylenol}
Answer 2: {Order a CT scan to check for brain tumors}
Preference Labeler Model
AI Preference
Label_0 = 0.6
Label_1 = 0.4
Dataset Creation
Chosen
Prompt: {What to prescribe for an acute minor headache?}
Answer 1: {Drink more water and take some tylenol}
Rejected
Prompt: {What to prescribe for an acute minor headache?}
Answer 2: {Order a CT scan to check for brain tumors}

Fig. 1. A diagram for the AI preference labeling process, with example prompt and responses.

highest risk in this context. To quantify this risk, we will leverage a customized version of shadow attacks, as originally proposed by Shokri et al. [7], with the full details outlined in the experimental section.

Formally, we define the membership inference risk as the probability $P(D_{private}|I_{external})$, which reflects the likelihood of inferring membership in $D_{private}$ based on $I_{external}$. The Membership Inference Attack Model can be defined as a function $A(x, I_{external})$, where x is a data point accessible through $I_{external}$. $A(x, I_{external}) \in \{0, 1\}$ predicts whether x belongs to $D_{private}$. We propose a modified version of shadow attacks to evaluate the effectiveness of membership inference attacks on the SFT model. The specific modifications will be elaborated on in the experimental section. Where our goal is to develop a privacy-preserving RLAIF pipeline that minimizes the membership inference risk, as measured by the attack success rate (ASR) calculated from $A(x, I_{external})$.

Training Effectiveness: We also aim to evaluate the effectiveness of RLAIF training on the model's performance, considering both accuracy and alignment with desired output characteristics. We use two different metrics to achieve this:

- **ROUGE-based Accuracy Assessment:** we denote the test dataset as D_{test}. Let M_{SFT}, M_{RLAIF}, and $M_{PP-RLAIF}$ represent the SFT model, vanilla RLAIF model, and privacy-preserving RLAIF model, respectively. We employ ROUGE metrics by Lin [22] to compare model outputs on D_{test} by calculating ROUGE scores R_{SFT}, R_{RLAIF}, and $R_{PP-RLAIF}$ for each model. and compare scores to assess relative accuracy improvements or degradations.
- **Pairwise Alignment Evaluation:** we define a pairwise alignment model $A(a1, a2)$ that compares two model outputs $a1$ and $a2$ and outputs a preference score indicating which response is better aligned with desired characteristics. This is done by conducting pairwise comparisons of generated responses using A. We then analyze preference scores to quantify model alignment. It is worth noting that this is similar to the work done by Lee *et al.* [3] in their comparisons of RLAIF models to other models.

Our investigation into these effectiveness measures acknowledges the potential for inherent, and potentially complex, trade-offs. Further research dedicated to understanding and optimizing these trade-offs emerges as a potential future direction.

IV. PriMa Algorithm

We introduce *PriMa*, an algorithm specifically tailored for data augmentation within RLAIF frameworks. The PriMa algorithm takes place between the SFT Model and the LLM model in the RLAIF pipeline as shown in figure 2. It operates by accordingly masking a designated percentage of tokens within a data point, comprising a prompt and its corresponding generated responses. This process is repeated iteratively to achieve the desired degree of obfuscation as illustrated in Figure 3.

The algorithm's primary objective is to establish a balance between two crucial requirements:

- **Privacy Preservation:** Masking tokens strategically prevents the ability of adversaries to infer sensitive information from the data, thereby enhancing privacy.
- **Content Preservation:** Maintaining sufficient coherence within the masked data is crucial to ensure its utility for the RLAIF labeler, as it relies on contextual understanding to provide meaningful feedback.

Key Parameters for granular control:

- $\rho\%$: This parameter governs the probability of token replacement within each iteration, directly influencing the extent of privacy preservation.
- N: This parameter dictates the number of iterations performed, incrementally augmenting the overall level of obfuscation.

The precise operational details of *PriMa* are encapsulated within Algorithm 4. *PriMa* offers a flexible and parameterized approach to privacy-preserving data augmentation in RLAIF, empowering researchers to calibrate the trade-off between pri-

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

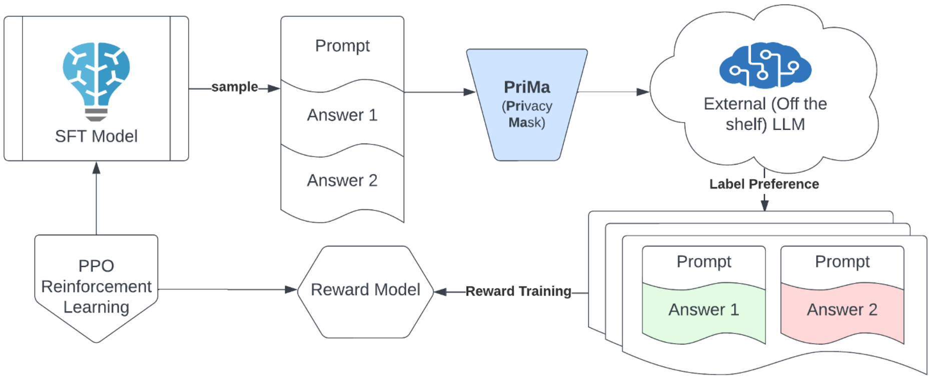

Fig. 2. The RLAIF pipeline with PriMa

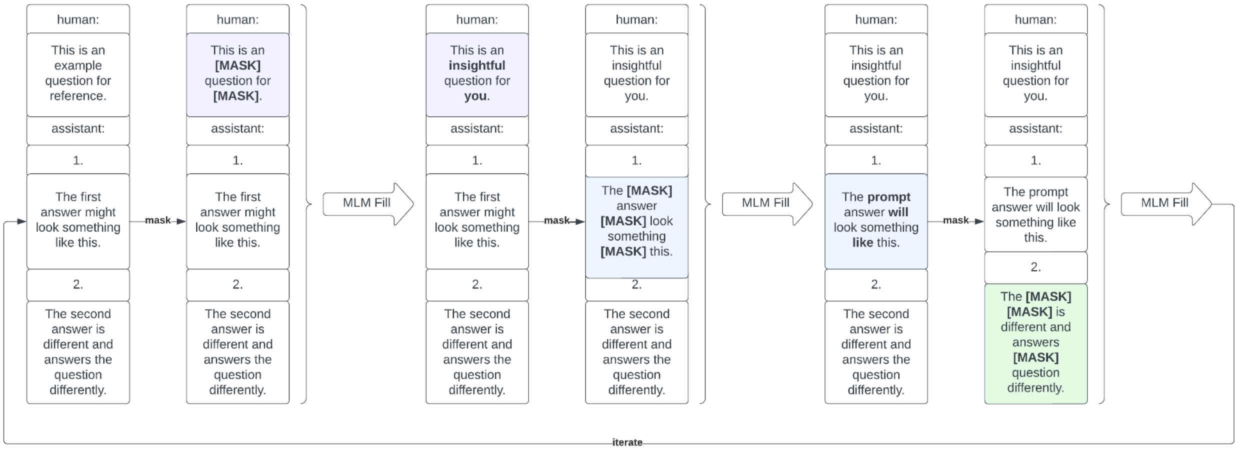
human:
This is an example question for reference.
assistant:
1.
The first answer might look something like this.
2.
The second answer is different and answers the question differently.
mask
human:
This is an [MASK] question for [MASK].
assistant:
1.
The first answer might look something like this.
2.
The second answer is different and answers the question differently.
MLM Fill
human:
This is an insightful question for you.
assistant:
1.
The first answer might look something like this.
2.
The second answer is different and answers the question differently.
mask
human:
This is an insightful question for you.
assistant:
1.
The [MASK] answer [MASK] look something [MASK] this.
2.
The second answer is different and answers the question differently.
MLM Fill
human:
This is an insightful question for you.
assistant:
1.
The prompt answer will look something like this.
2.
The second answer is different and answers the question differently.
mask
human:
This is an insightful question for you.
assistant:
1.
The prompt answer will look something like this.
2.
The [MASK] [MASK] is different and answers [MASK] question differently.
MLM Fill
iterate

Fig. 3. A simplified step-by-step of PriMa.

vacy and content preservation to align with specific application requirements.

A. Probabilistic Mask and Replacement

To safeguard the integrity of both semantic content and privacy within the *PriMa* algorithm, we execute a multi-stage process that preserves the underlying structure of the input data:

1) **Sectional Disassembly:** The input data is parsed into its constituent sections—the prompt, Answer 1, and Answer 2—to ensure independent masking while maintaining their interrelationships.
2) **Intra-Section Masking with Probabilistic Token Replacement:** Within each section, we iterate through each token and probabilistically supplant $\rho\%$ of them with the [MASK] token, strategically obfuscating sensitive information while preserving contextual cues.

Fig. 4. PriMa Algorithm

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

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

3) **Context-Aware Token Prediction via TinyBERT:** The masked sections are subsequently fused and collectively presented to a TinyBERT model [23], a compact and efficient Transformer-based language model. This model leverages its contextual understanding to predict the most plausible replacements for the [MASK] tokens, ensuring semantic coherence within the masked text.
4) **Reassembly and Iteration:** The masked tokens within each section are then substituted with their corresponding TinyBERT-predicted counterparts, and the sections are reassembled to reconstitute the complete data point. This process is reiterated for the remaining sections (Answer 1 and Answer 2), fostering a comprehensive obfuscation strategy.

Separating the text into sections and masking them separately gives several advantages as follows:

- **Enhanced Labeling Accuracy:** By preserving structural integrity and coherence, we facilitate the external LLM's ability to accurately comprehend and label the masked data, yielding high-quality training material for the reward model.
- **Fortified Privacy:** The granular, probabilistic token masking technique impedes potential adversaries from reverse-engineering the algorithm and effectively reconstructing the original text, maintaining privacy protection.
- **Balanced Trade-off:** The adjustable parameters $\rho\%$ and N empower researchers to precisely calibrate the degree of privacy preservation and content coherence, striking an optimal balance that aligns with the specific requirements of the RLAIF pipeline.

B. Iterative Probabilistic Masking: Balancing Privacy and Coherence

PriMa's core masking mechanism, probabilistic token replacement, allows for iteratively enhancing privacy while navigating the delicate balance with content coherence. By repeating the masking process N times, we progressively replace more tokens within each data point. Notably, maintaining stability throughout these iterations is crucial. This is achieved by ensuring only a fraction of the tokens are masked at any given time, preventing drastic alterations to the data structure. The privacy benefits, however, increase with each iteration. As more tokens are probabilistically replaced, the original data becomes increasingly obscured.

V. Dataset and Base Models

This section lays the groundwork for the extensive experimental study, which forms the backbone of our evaluation of the *PriMa* algorithm. Our objective necessitates training multiple RLAIF models under different configurations, requiring a constructed and split dataset.

A. Dataset Splitting

We utilize the MedQuAD dataset [24], a publicly available repository containing 16.4K medical question-answer pairs. This rich resource provides a realistic and relevant domain for evaluating the effectiveness of our approach. The dataset is partitioned into distinct subsections to facilitate various aspects of the experiment:

1) **Fine-tuning:** 5.74K question-answer pairs are allocated for fine-tuning both the target SFT model and the shadow model used for membership inference attacks. This ensures both models are well-equipped for their respective tasks.
2) **Vanilla RLAIF Pipeline:** 820 out-of-sample questions serve as the input for the standard RLAIF pipeline, allowing us to compare its performance to models utilizing *PriMa* augmented data.
3) **Shadow Classification:** An additional 820 out-of-sample questions are reserved for the shadow classification task, enabling us to assess the effectiveness of membership inference attacks on different data configurations.
4) **Reinforcement Learning:** Dedicated sets of 1.64K questions each are designated for reinforcement learning within the RLAIF pipeline and for testing the final models. This dedicated resource ensures robust training and accurate performance evaluation.

This data-splitting strategy creates a rigorous evaluation framework, enabling us to compare the performance of models trained with and without *PriMa* across various tasks. Each partition plays a crucial role in revealing the efficacy of our proposed approach in terms of privacy preservation and model alignment.

B. Mock Preference Labeler Model

Given the resource constraints and desire for a controlled environment, we opted for a synthetic approach to preference labeling, rather than directly employing a commercial LLM in the experimental phase. This involved fine-tuning a DistilBERT model [25] for text sequence classification based on the hh-rlhf dataset [26] curated by Anthropic. This dataset specifically focuses on the ”helpfulness” aspect of model responses [26], a crucial criterion for our evaluation.

To leverage the hh-rlhf data effectively, we transformed it into a binary classification task. For each prompt and its two corresponding answers, the model predicts ”0” if the first answer is deemed more helpful and ”1” if the second is preferred. This simplified formulation facilitates efficient and consistent labeling within the controlled setting. This preference labeler model plays a pivotal role in various stages of our evaluation:

1) **Reward Model Training:** We utilize the model as a surrogate for the ”External LLM” depicted in Figure 2, allowing us to generate labeled training data for both the Vanilla Reward Model and the Privacy Preserving Reward Model. This enables a controlled comparison of their performance under different data configurations.
2) **Alignment Assessment:** In the ”Alignment Increase” section, we reuse the model to act as a grader, determining the win rates of different models when facing head-to-head comparisons on test prompts. This provides

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

valuable insights into the impact of our *PriMa* algorithm on model alignment with desired characteristics.

By employing a well-chosen and carefully constructed mock preference labeler model, we achieve both resource efficiency and a controlled environment for evaluating the effectiveness of our *PriMa* algorithm. This approach offers a reliable foundation for assessing its impact on privacy preservation and model alignment within the RLAIF pipeline. The model is available at [27].

C. Target Supervised Fine-Tuned Model (Target SFT)

This section introduces the Target Supervised Fine-Tuned Model (Target SFT), which serves as a foundational element in our evaluation of the *PriMa* algorithm within the RLAIF framework. We leverage the well-established T5-small model [28] pre-trained by Google AI, fine-tuned on a dedicated portion of the MedQuAD dataset [24]. This dataset offers a rich resource of medical question-answer pairs sourced from diverse sources, making it highly relevant to the domain of healthcare information access. Importantly, we intentionally hold back some parts of the MedQuAD dataset for subsequent stages of the experiment, maximizing data utilization and preventing potential overfitting issues.

The Target SFT fulfills two crucial roles within our research:

- **RLAIF Baseline:** This model serves as the initial starting point for the RLAIF training process. By fine-tuning its parameters on the chosen MedQuAD data, we establish a baseline performance against which the RLAIF models incorporating *PriMa*-augmented data can be compared. This comparison allows us to assess the efficacy of our proposed approach in terms of model alignment and accuracy within the RLAIF pipeline.
- **Privacy Evaluation Target:** The Target SFT also plays a vital role in our privacy evaluations, specifically concerning membership inference attacks. Its availability enables us to analyze the effectiveness of *PriMa* in obfuscating sensitive information within training prompts, thereby mitigating the risk of unauthorized data inference.

The model is available at [29].

D. Vanilla RLAIF Pipeline

This section dives into the Vanilla RLAIF Pipeline, which serves as the baseline against which we evaluate the effectiveness of our *PriMa* algorithm within the RLAIF framework.

1) Vanilla Reward Model (Vanilla RM): The main ingredient of the RLAIF pipeline is the Vanilla Reward Model (Vanilla RM), trained on a carefully constructed dataset designed to represent a realistic training scenario with potential privacy concerns. We achieve this by combining several key elements:

- **Balanced Data:**
  - Response Pairs: We extract pairs of responses (generated by the Target SFT) to questions from another portion of the MedQuAD dataset. This ensures the responses are relevant to the domain but distinct from the training data used for the Target SFT, mitigating potential overfitting issues.
  - Randomly Sampled In-sample Prompts: An equal number of randomly sampled prompts from the Target SFT's training data are added. This introduces realistic diversity and reflects the unknown proportion of sensitive information in real-world training data.
- **Varied Temperatures:** To encourage diversity in responses, the Target SFT generates responses with varied temperatures during the response pair extraction process. This further strengthens the generalizability of the trained model.
- **Mock Preference Labeler Integration:** Questions and paired responses are formatted and presented to the previously mentioned mock preference labeler model. This model evaluates the ”helpfulness” of each response, providing valuable labels for training the Vanilla RM.
- **DistilRoBERTa Fine-tuning:** We leverage a DistilRoBERTa model [25], [30] as the base architecture for the Vanilla RM. This model is fine-tuned for single-label text classification, generating a single score for each response, which later serves as the reinforcement learning signal within the RLAIF pipeline.

This reward model is available at [31]

2) Vanilla RLAIF Model: Using the Vanilla RM from the previous section, we utilize the Proximal Policy Optimization (PPO) Algorithm proposed by Schulman *et al.* [32] 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.

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 [33].

E. Privacy Preserving RLAIF Pipeline

We begin a separate process of privacy-preserving RLAIF as shown in Figure 2. This mirrors the Vanilla RLAIF pipeline with the addition of our novel *PriMa* algorithm.

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 data points into “chosen” and “rejected” using the same process as the Vanilla RM.

This dataset now contains “masked” responses that no longer directly resemble the original outputs of the Target SFT, but remain 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

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

text classification, using the same training parameters as the Vanilla RM to maintain comparability. This model is available at [34].

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 [35].

F. Shadow Model for Membership Inference Attack

We propose a novel approach to evaluate the privacy preservation efficacy of the PriMa algorithm. Inspired by the framework established by Shokri et al. [7], we focus on the data labeling stage within the RLAIF pipeline as a critical point for potential privacy leakage. To the best of our knowledge, this constitutes the first dedicated investigation into privacy vulnerabilities within this specific stage, precluding direct comparisons with existing baselines.

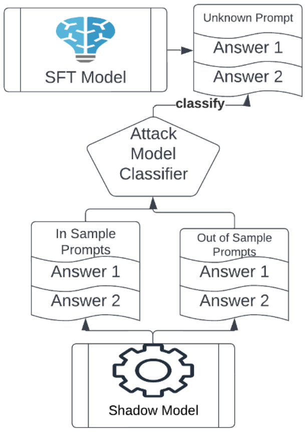

Fig. 5. The Shadow model for membership inference attack diagram

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 presented in Figure 5. In practice, there are varying degrees of success with this strategy since the attack depends completely on the date chosen, but the primary measure is the precision of the attack [7].

The first step in implementing this attack is to train a model that 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 with 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 that are publicly available. This shadow model is available at [36]. 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 at 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 data point 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.

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 [37]. 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.

VI. *PriMa's* EFFICACY

The study design involves comprehensively controlling various parameters of the *PriMa* algorithm, such as the masking probability ($\rho\%$) and the number of iterations (N). By systematically comparing the performance of models trained on data subject to different *PriMa* settings, we can gather valuable insights into the algorithm's efficacy. Notably, this comparative analysis will address two key aspects: Privacy Preservation, for which we will evaluate the degree to which *PriMa* successfully obscures sensitive information in the input data, and RLAIF Performance, for which we will measure the impact of *PriMa*-augmented data on the performance of RLAIF models. This entails examining metrics like accuracy, alignment with desired outputs, and overall effectiveness in guiding reinforcement learning. By carefully navigating the intricate balance between privacy preservation and RLAIF effectiveness, we can establish the optimal configuration for *PriMa* within the context of specific RLAIF applications.

A. Privacy Preservation

In the domain of evaluating privacy preservation, precision takes center stage, quantifying the effectiveness of membership inference attacks. This metric shines a light on the percentage of in-sample data points that the attack model successfully identifies.

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

TABLE I
Membership Inference Attack Results

Prima Augmentation %	Attack Precision %
0% (Base Data)	50.19
30%	49.37
40%	48.99
50%	48.56

TABLE II
ROUGE scores for each model on the test dataset. Vanilla RLAIF involved no masking, while the three rightmost columns represent Privacy RLAIF with three different masking probabilities

	SFT	Vanilla RLAIF	0.3	0.4	0.5
ROUGE-1	26.79	26.01	22.77	20.40	22.30
ROUGE-2	11.95	12.11	12.49	11.12	12.05
ROUGE-L	22.09	21.48	19.32	17.51	19.09

As Table I reveals, the starting point—the Vanilla RLAIF pipeline's ”base data”—exhibits a surprisingly low attack precision. This observation can be attributed to two primary factors:

- **Complex Classification Landscape:** The inherent intricacies of the classification problem itself may pose inherent challenges for the attack model, hindering its ability to accurately discern membership.
- **Data Scarcity:** The size of the dataset utilized to train the attack model might also play a role. Limited data availability can constrain the model's learning capacity, potentially leading to underfitting and underwhelming attack performance.

However, when the data is filtered through the *PriMa* algorithm, a subtle but noticeable decrease in precision can be noted. This reduction in attack success suggests that *PriMa* effectively camouflages a portion of the in-sample data points, Obfuscating sensitive information through a layer of protection, ensuring its confidentiality.

B. RLAIF Performance: ROUGE Stability

Our primary objective is to ensure that the enhanced privacy afforded by the *PriMa* algorithm does not come at the expense of language model performance. To comprehensively evaluate the quality of responses generated by the various models, we employed a battery of ROUGE metrics [22] based on comparisons with curated reference responses in the test dataset.

1) ROUGE Score Analysis: Table II demonstrates that most ROUGE metrics remain relatively stable across all three models (SFT, Vanilla RLAIF, and Privacy-Preserving RLAIF). However, a notable decrease in the ROUGE-1 score for the Privacy-Preserving RLAIF model is observed. This was an anticipated consequence of incorporating token masking and replacement within the reward model, as it inevitably alters the model's vocabulary distribution. Interestingly, 0.4 masking rate resulted in lower scores than 0.3 and 0.5.

Moving beyond ROUGE-1, the analysis reveals other informative patterns. The ROUGE-2 score, which assesses two-word phrase similarity, indicates that both RLAIF models can maintain coherence in terms of generating bi-grams comparable to the reference text, exceeding the performance of the SFT model in this aspect. This observation mitigates concerns raised by the reduced ROUGE-1 score and reinforces the notion that the RLAIF models can deliver accurate answers on par with the SFT model.

TABLE III
Results from head-to-head annotations by the AI Preference Labeler

	SFT	Vanilla RLAIF
Vanilla RLAIF	58.65%	N/A
PriMa 0.3	66.40%	56.65%
PriMa 0.4	70.42%	61.89%
PriMa 0.5	67.13%	58.29%

Finally, the ROUGE-L score, capturing overall structural similarity, suggests that both RLAIF models generate responses with slightly less structural resemblance to the reference text compared to the SFT model. While this divergence may be attributed to the intrinsic features of the Proximal Policy Optimization (PPO) algorithm employed in the RLAIF training process, it does not definitively imply lower response quality. Further investigation and analysis are necessary to fully understand the implications of this observation, along with the apparent dip in overall ROUGE score around 0.4 masking rate.

C. Alignment Increase

Given that the benefit of RLAIF's advantage lies in its ability to align generated responses with desired characteristics, evaluating the potential effects of our novel *PriMa* algorithm on this crucial metric is important.

To achieve this, we leverage the simulated preference labeler model (introduced in Section V-B) to conduct pairwise comparisons of model responses within the test dataset. This enables the calculation of win rates, whereby a ”win” signifies a response that is deemed more aligned and, in the context of this study, more ”helpful” than its counterpart.

As illustrated in Table III, both RLAIF models emerge victorious against the SFT model in a majority of comparisons, echoing the findings of Lee et al. [3], albeit to a slightly lesser extent.

Crucially, the model subjected to the privacy-preserving RLAIF pipeline, incorporating the *PriMa* algorithm, exhibits an even more pronounced win rate against the SFT model compared to the vanilla RLAIF approach. Moreover, when directly compared against the vanilla RLAIF model, the preference labeler demonstrates a preference for the responses generated by our *PriMa*-enhanced model. In contrast to the previous experiment, a masking rate of 0.4 had higher performance in head to head trials than both 0.3 and 0.5. That cause of this is currently unclear, but may have to do with the proportion of complex phrases that get replaced resulting in easier alignment up to a point, but lowering similarity with the original phrase.

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.

Based on the observed alignment improvement, we hypothesize that the *PriMa* algorithm might act as a noise filter for the preference labeler. This potentially results in higher-quality annotations, leading to more effective training of the reward model. Expanding on this hypothesis and exploring its generalizability to broader RLAIF contexts represents a significant future research opportunity.

VII. Conclusion

PriMa stands as a testament to the possibility of harmonizing privacy and performance within the RLAIF domain. By offering a sophisticated approach to data preprocessing, the algorithm paves the way for building and deploying RLAIF models that deliver accurate, aligned responses while safeguarding sensitive information, ultimately propelling responsible innovation in the realm of language models. 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.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[2] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[3] H. Lee, S. Phatale, H. Mansoor, K. Lu, T. Mesnard, C. Bishop, V. Carbune, and A. Rastogi, "Rlaif: Scaling reinforcement learning from human feedback with ai feedback," *arXiv preprint arXiv:2309.00267*, 2023.

[4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback, 2022," *URL https://arxiv.org/abs/2203.02155*, vol. 13, 2022.

[5] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[8] X. Pan, M. Zhang, S. Ji, and M. Yang, "Privacy risks of general-purpose language models," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 1314–1331.

[9] F. Mireshghallah, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. Shokri, "Quantifying privacy risks of masked language models using membership inference attacks," *arXiv preprint arXiv:2203.03929*, 2022.

[10] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh, "Propile: Probing privacy leakage in large language models," *arXiv preprint arXiv:2307.01881*, 2023.

[11] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.

[12] N. Kandpal, K. Pillutla, A. Oprea, P. Kairouz, C. A. Choquette-Choo, and Z. Xu, "User inference attacks on large language models," *arXiv preprint arXiv:2310.09266*, 2023.

[13] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin, "Analyzing leakage of personally identifiable information in language models," *arXiv preprint arXiv:2302.00539*, 2023.

[14] A. Jagannatha, B. P. S. Rawat, and H. Yu, "Membership inference attack susceptibility of clinical language models," *arXiv preprint arXiv:2104.08305*, 2021.

[15] M. Duan, A. Suri, N. Mireshghallah, S. Min, W. Shi, L. Zettlemoyer, Y. Tsvetkov, Y. Choi, D. Evans, and H. Hajishirzi, "Do membership inference attacks work on large language models?" 2024. [Online]. Available: https://arxiv.org/abs/2402.07841

[16] C. Patsakis and N. Lykousas, "Man vs the machine: The struggle for effective text anonymisation in the age of large language models," *arXiv preprint arXiv:2303.12429*, 2023.

[17] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 596–606, 2017.

[18] S. Murthy, A. A. Bakar, F. A. Rahim, and R. Ramli, "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, 2019, pp. 306–309.

[19] N. Zhou, Q. Wu, Z. Wu, S. Marino, and I. D. Dinov, "Datasiftertext: Partially synthetic text generation for sensitive clinical notes," *Journal of Medical Systems*, vol. 46, no. 12, p. 96, 2022.

[20] X. Yue, H. A. Inan, X. Li, G. Kumar, J. McAnallen, H. Shajari, H. Sun, D. Levitan, and R. Sim, "Synthetic text generation with differential privacy: A simple and practical recipe," *arXiv preprint arXiv:2210.14348*, 2022.

[21] S. Kweon, J. Kim, J. Kim, S. Im, E. Cho, S. Bae, J. Oh, G. Lee, J. H. Moon, S. C. You *et al.*, "Publicly shareable clinical large language model built on synthetic clinical notes," *arXiv preprint arXiv:2309.00237*, 2023.

[22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[23] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[24] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–23, 2019.

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[26] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[27] D. Gao, "hh-labeler model," https://huggingface.co/davidgaofc/hh-labeler, 2024.

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[29] D. Gao, "Sft_med_t model," https://huggingface.co/davidgaofc/SFT_Med_t, 2024.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[31] D. Gao, "Rm_base model," https://huggingface.co/davidgaofc/RM_base, 2024.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[33] D. Gao, "Ppo_base model," https://huggingface.co/davidgaofc/PPO_base, 2024.

[34] ——, "Rm_prima model," https://huggingface.co/davidgaofc/RM_prima, 2024.

[35] ——, "Ppo_prima model," https://huggingface.co/davidgaofc/PPO_prima, 2024.

[36] ——, "Sft_shadow model," https://huggingface.co/davidgaofc/SFT_shadow, 2024.

[37] ——, "Shadowattackf model," https://huggingface.co/davidgaofc/ShadowAttackF, 2024.

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on July 25,2025 at 15:14:32 UTC from IEEE Xplore. Restrictions apply.