Privacy Preserving RLAIF using Masking Algorithms

David Gao

Outline

1. Introduction
2. RLAIF Background
3. Experimental Setup
4. Vanilla RLAIF Pipeline
5. Proposed Algorithm
6. Privacy Preserving RLAIF Pipeline
7. Attack Model
8. Model Evaluations
9. Discussion
10. Demo (if we have extra time)

Introduction

- Motivation
- Goals of the solution

Motivation

- RLAIF has shown to improve model hallucinations among many other positive qualities for large language models.
- This can be utilized to align smaller models, making them perform almost as well as the larger ones.
- If we can preserve privacy in the RLAIF pipeline, we can utilize this in fields where organizations where privacy is very important like medicine and finance.

Goals

<u>Privacy Preservation</u>

- Protecting against membership inference attacks
- Precision of shadow attack

<u>Effectiveness of RLAIF</u>

- ROUGE stability
- Alignment Improvement

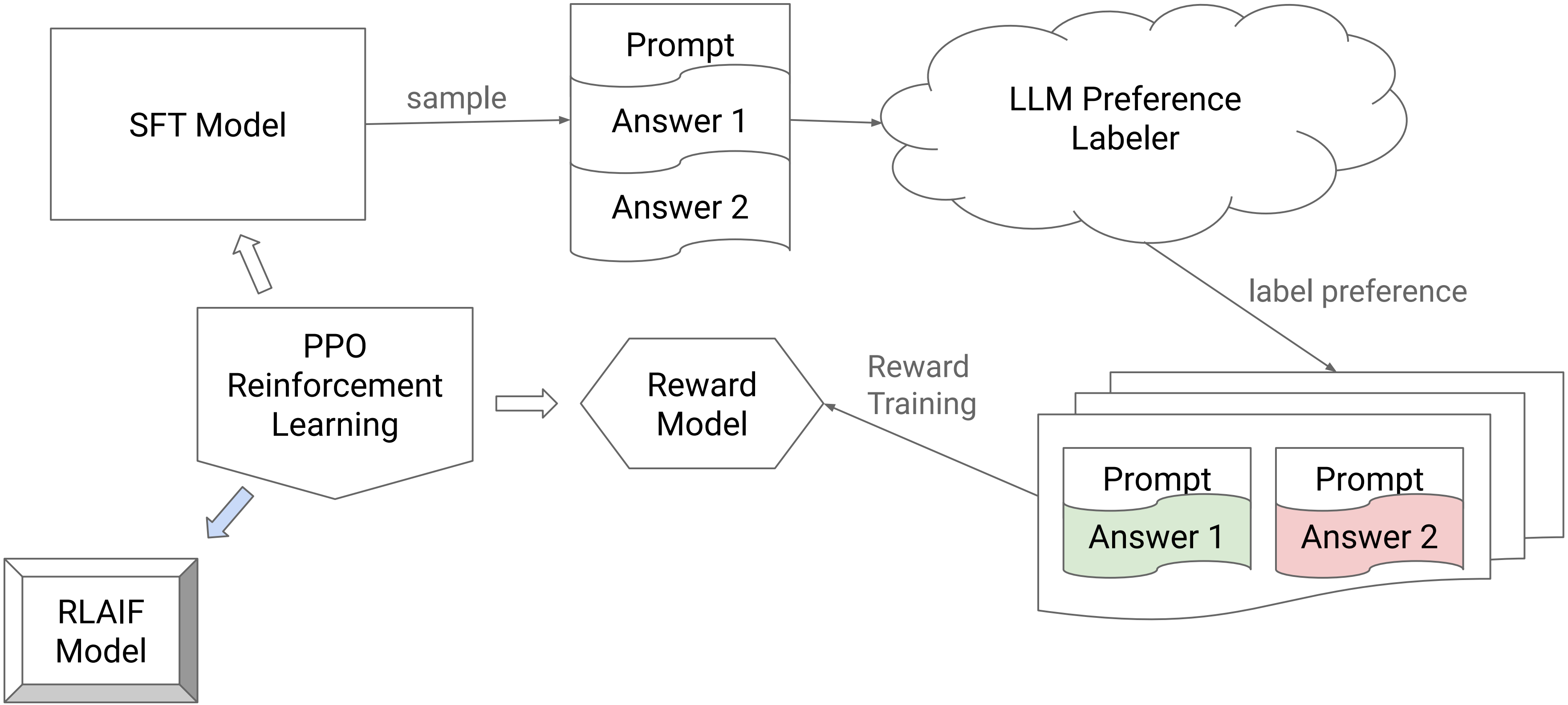
RLAIF Background

- What is RLAIF in the context of Large Language Models?
- What does the training and alignment pipeline look like?

Reinforcement Learning from AI Feedback

- Proposed way of scaling up Reinforcement Learning from Human Feedback
- Used for Alignment of Large Language Models

RLAIF Pipeline



Experimental Setup

- Dataset
- Base Model (SFT)
- AI Preference Labeler Model

MedQuAD Dataset

16.4K Question and Answer pairs collected by Ben *et al.*

Split:

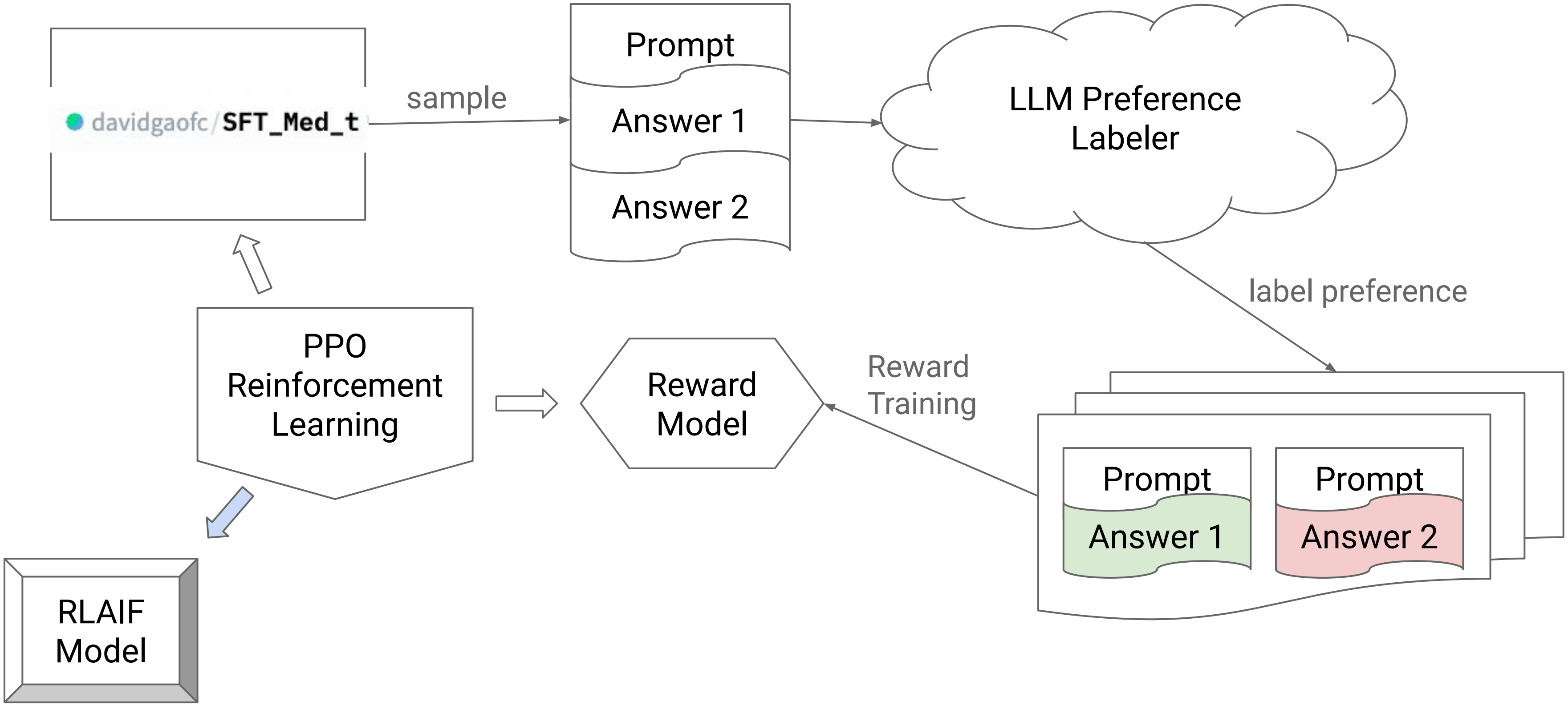
- 5.74K Target SFT Training set
- 5.74K Shadow SFT Training set
- 820 Out-of-sample questions for Vanilla RLAIF pipeline.
- 820 out-of-sample questions for shadow classification task.
- 1.64K questions designated for Reinforcement Learning.
- 1.64K Testing set.

Base SFT

davidgaofc/SFT_Med_t

- T5-Small
- Fine-tuned on 5.74K SFT Training split
- Sequence to Sequence

RLAIF Pipeline



AI Preference Labeler Model

davidgaofc/hh-labeler

DistilBERT model fine-tuned on the Anthropic/hh-rlhf dataset.

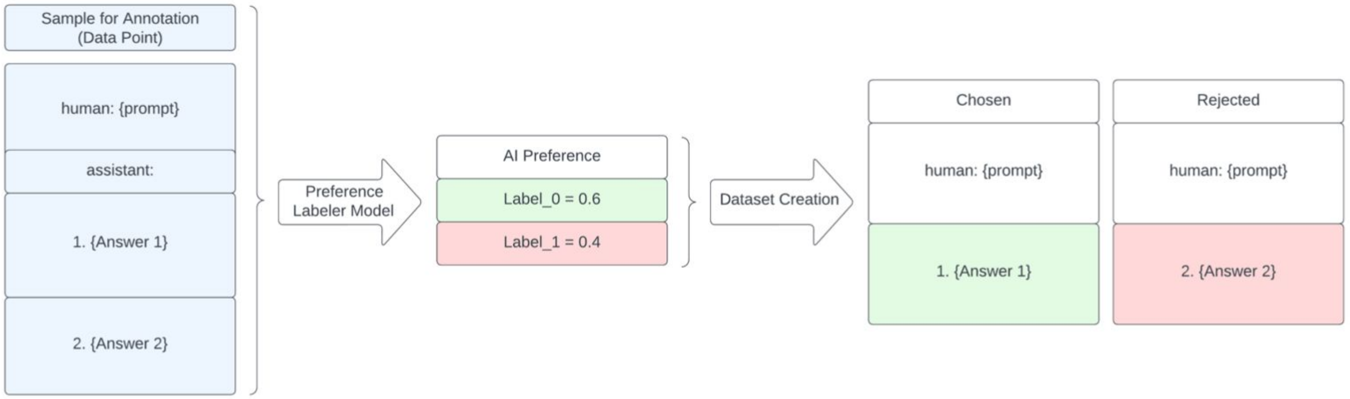
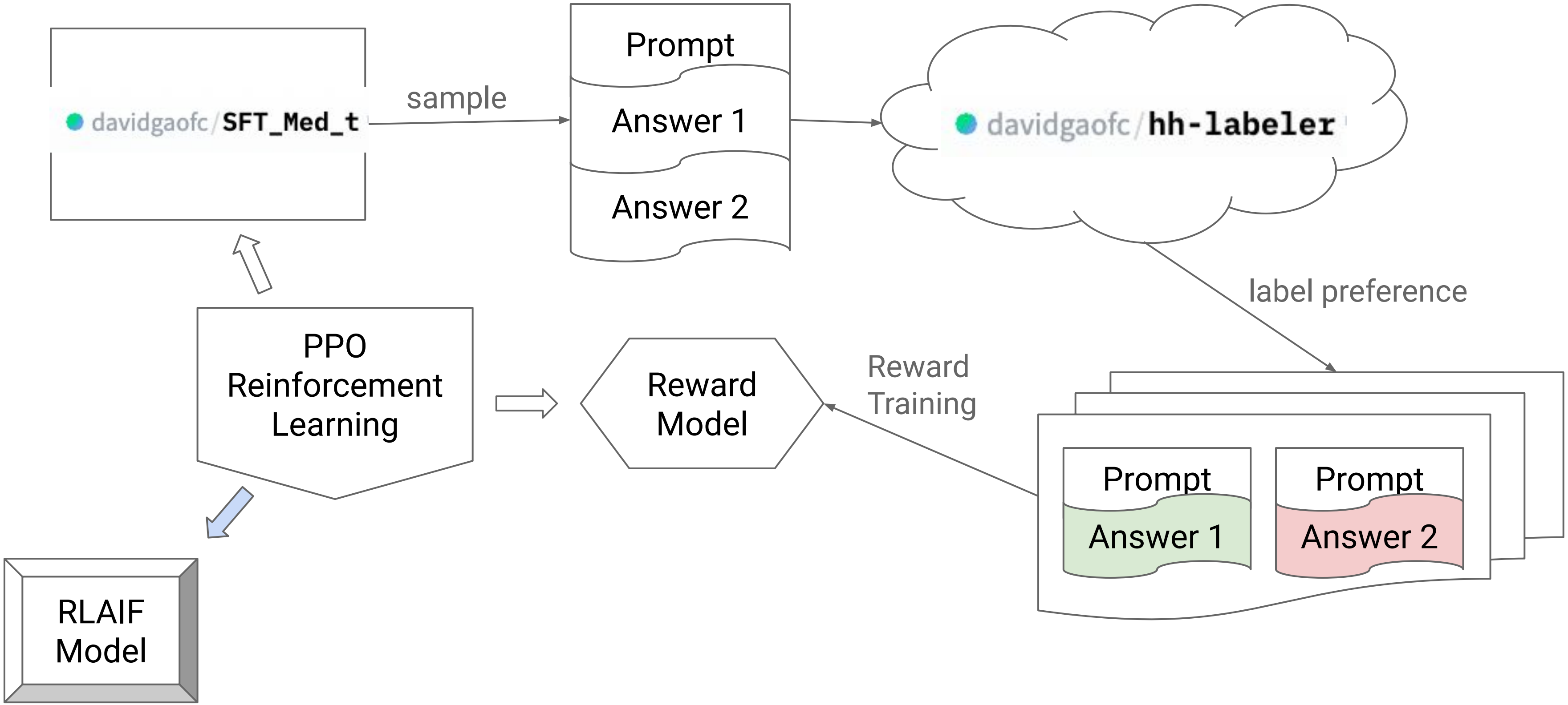


Figure 3. A diagram for the AI preference labeling process.

RLAIF Pipeline



Vanilla RLAIF Pipeline

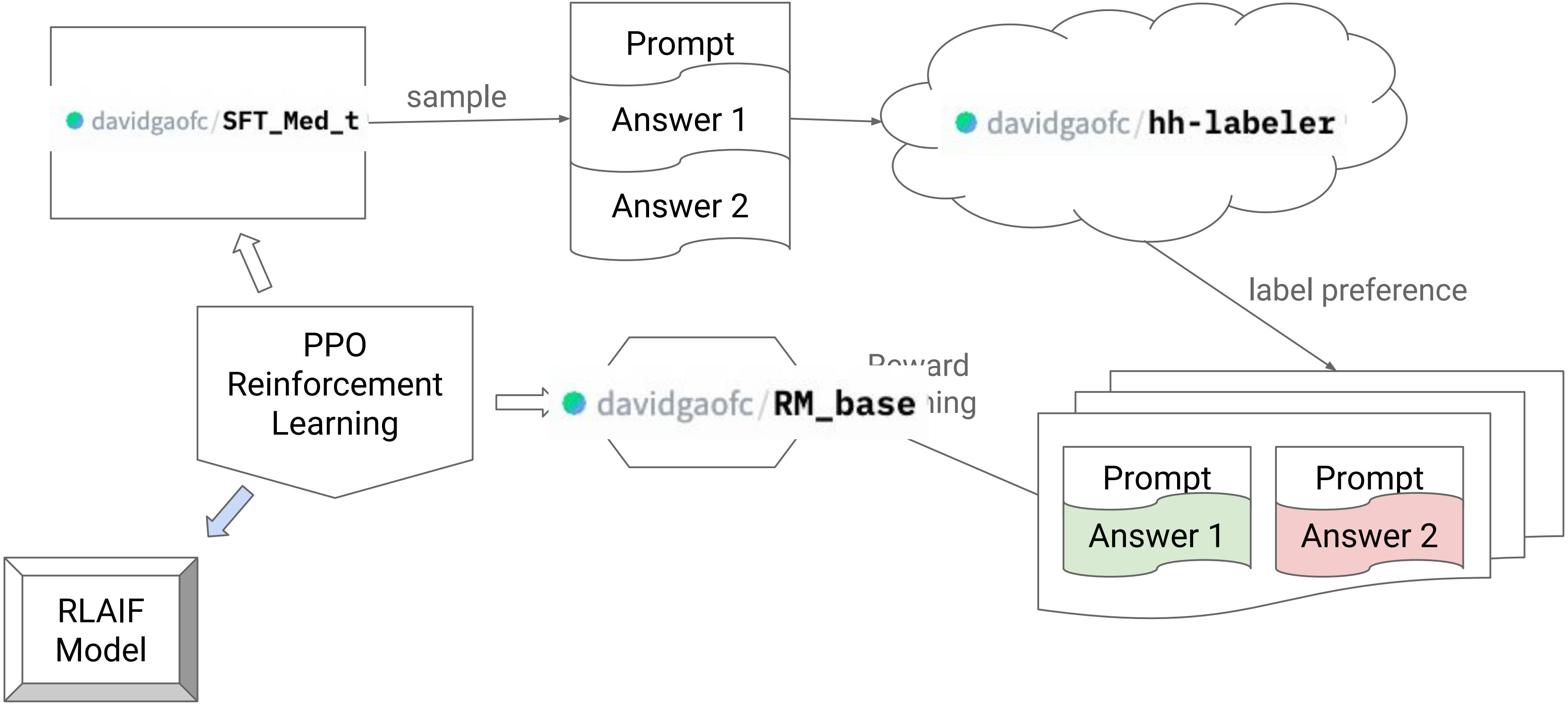
- Training the Reward Model
- Aligning the Vanilla RLAIF Model

Reward Model

davidgaofc/RM_base

- DistilRoBERTa
- Fine-tuned on labeled dataset (a mix of in sample and out of sample data points from the SFT)
- Text classification (single label)

RLAIF Pipeline

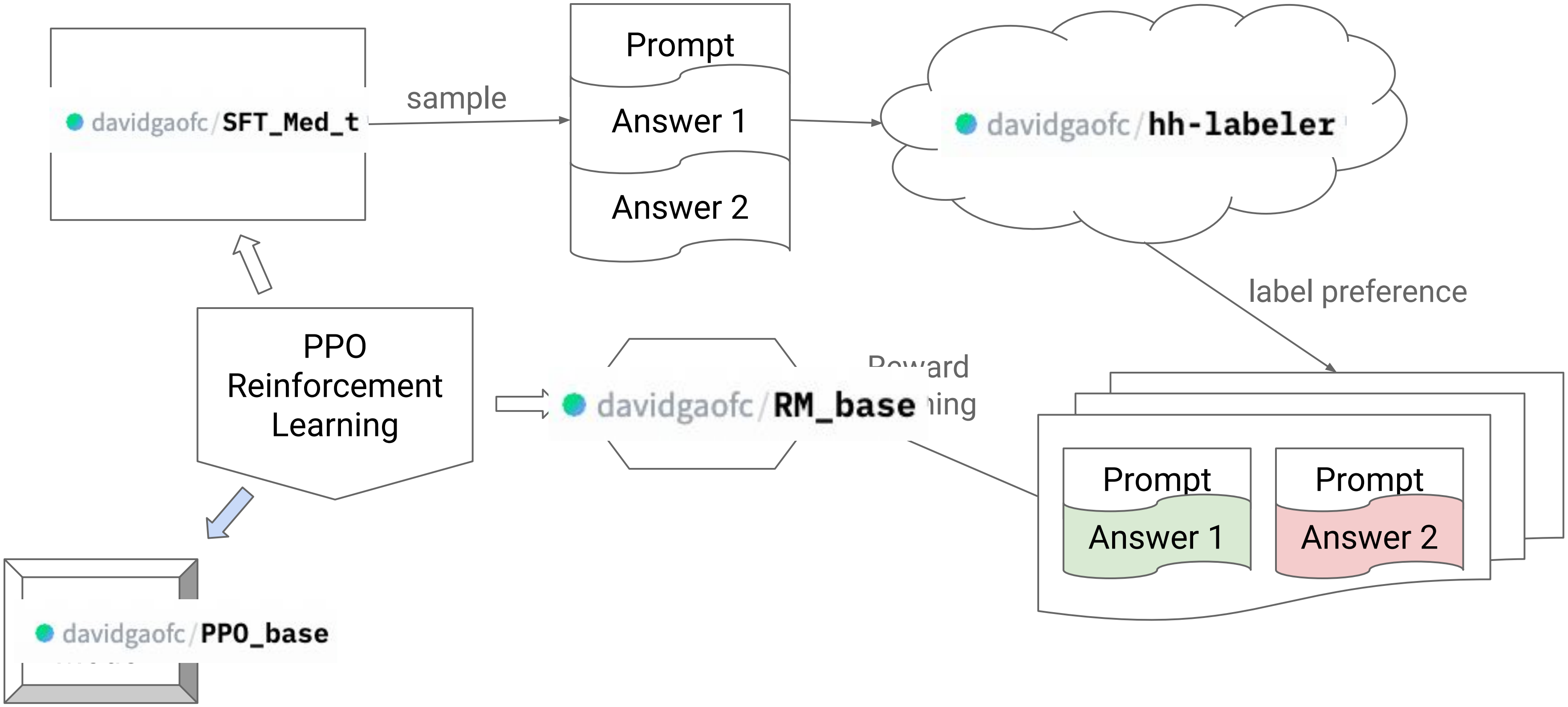


davidgaofc/PPO_base

Vanilla RLAIF

- Utilize the Proximal Policy Optimization Algorithm proposed by Schulman *et al.*
- Use Reward Model as RL signal
- Train using subsection of dataset designated for RL

RLAIF Pipeline



Proposed Algorithm (**Privacy Mask** - PriMa)

- General Architecture
- Practical Tradeoffs

Architecture

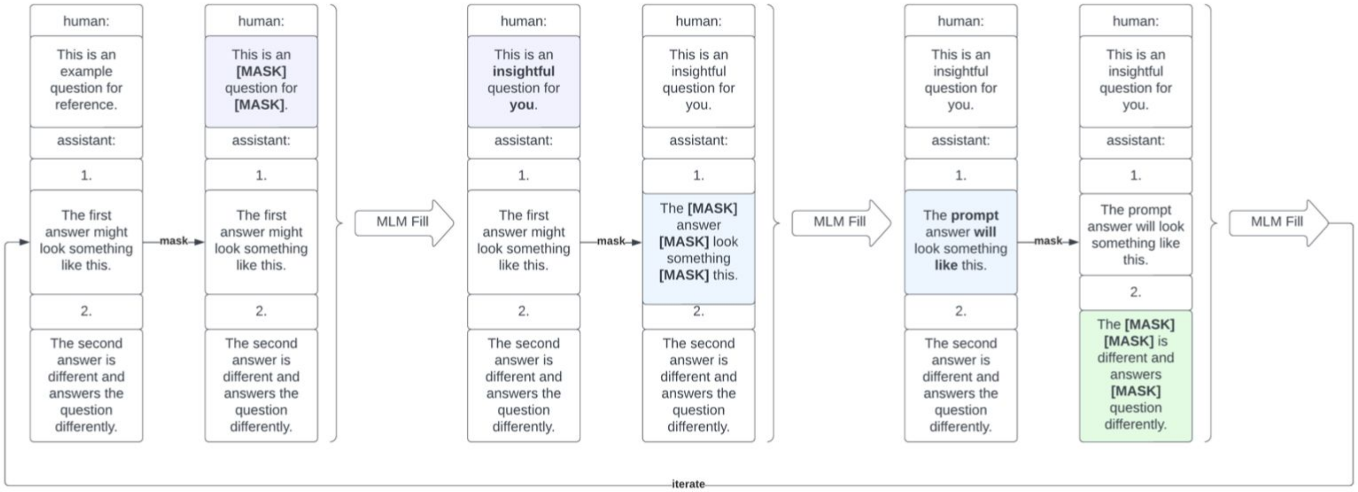


Figure 2. A simplified step by step of PriMa.

Pseudocode

Algorithm 1 PriMa Algorithm

```
Require: input, iterations, proportion
for i = 1 to iterations do
    prompt, answer1, answer2 ← Split(input)
    sections ← {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)
```

Trade Offs

Privacy

- If we mask with higher probability, more tokens will be replaced. (More obfuscation)

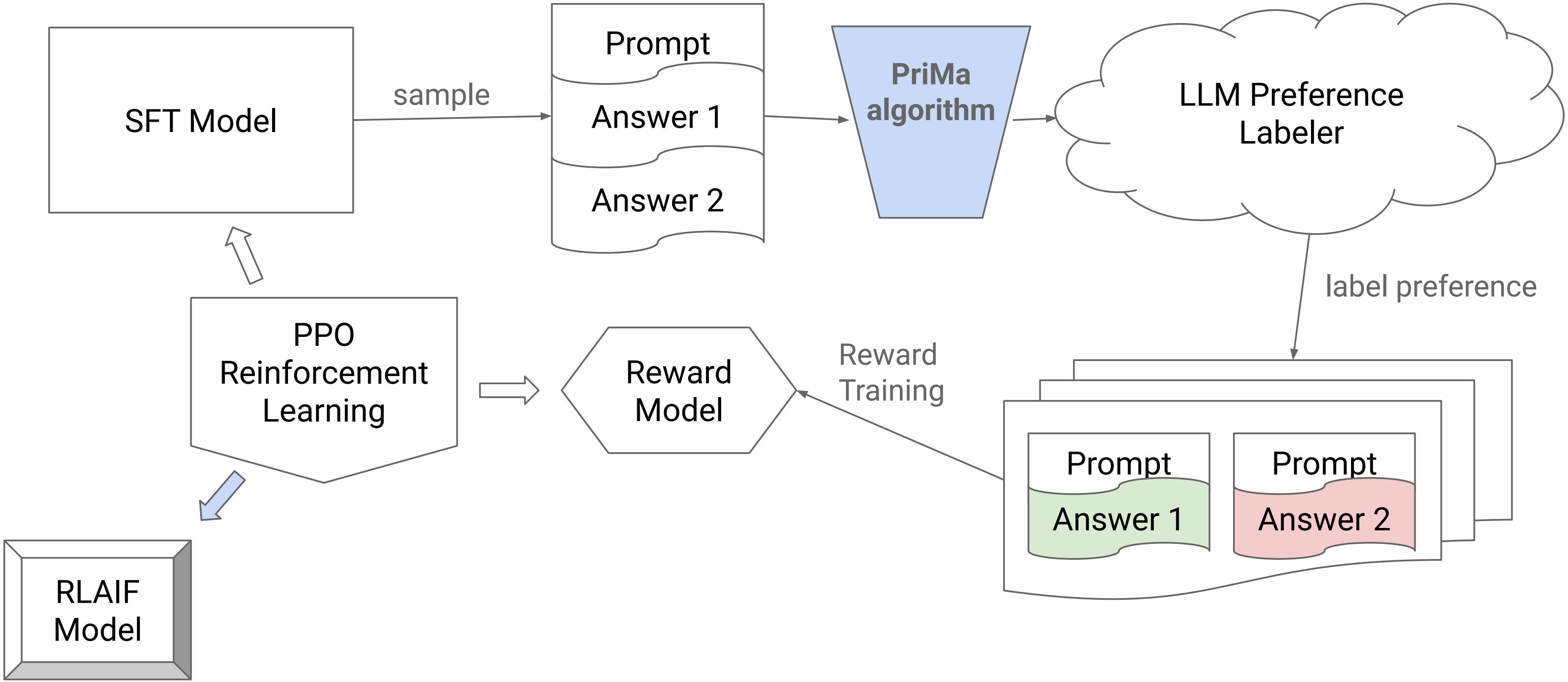
Coherence

- If we replace more tokens, there is a higher likelihood that the data points will lose coherence. (Noisier data for preference labeler)

Privacy Preserving RLAIF Pipeline (using PriMa)

- Data augmentation
- Training the PriMa Reward Model
- Aligning the PriMa RLAIF Model

PriMa Augmentation

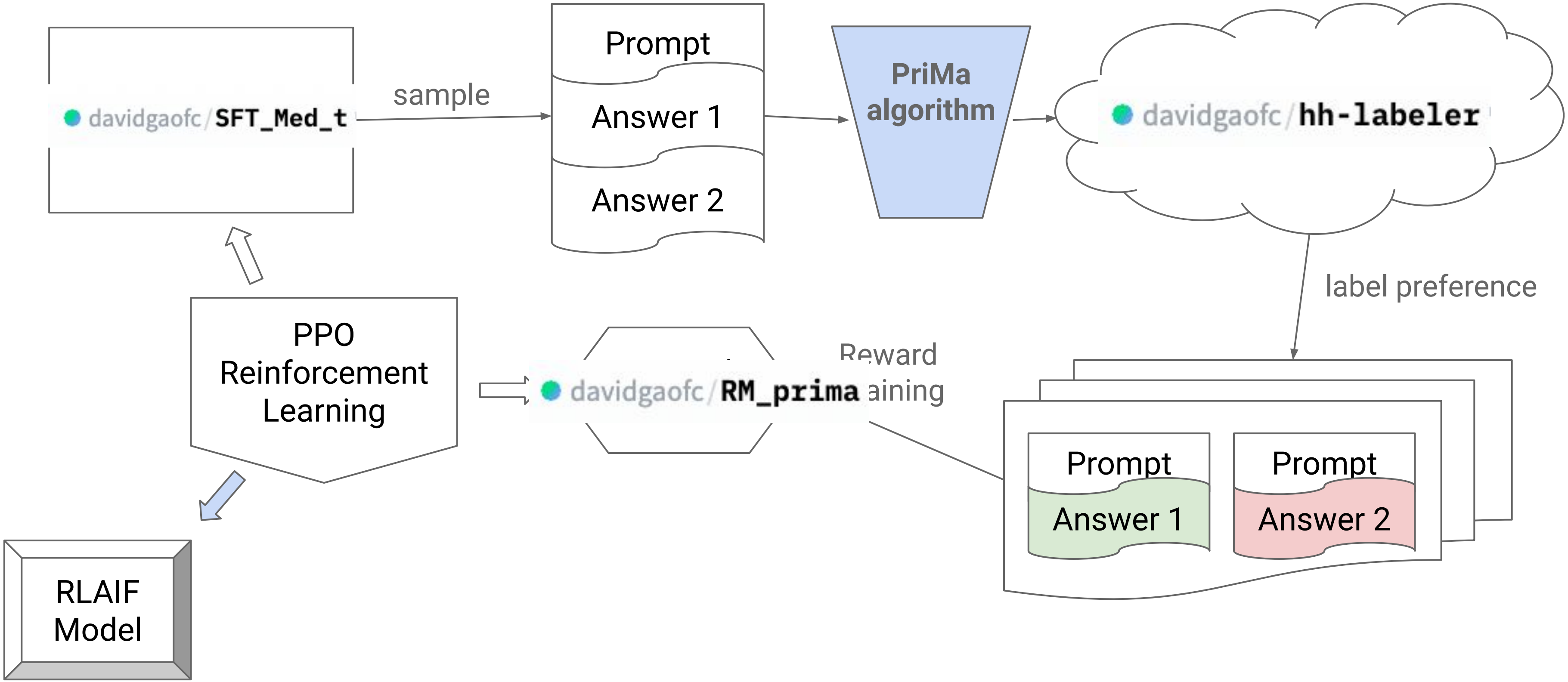


Privacy Preserving Reward Model

davidgaofc/RM_prima

- DistilRoBERTa
- Fine-tuned on the original RM dataset which is passed through the PriMa algorithm and relabeled.
- Text classification (single label)

Privacy Preserving RLAIF Pipeline

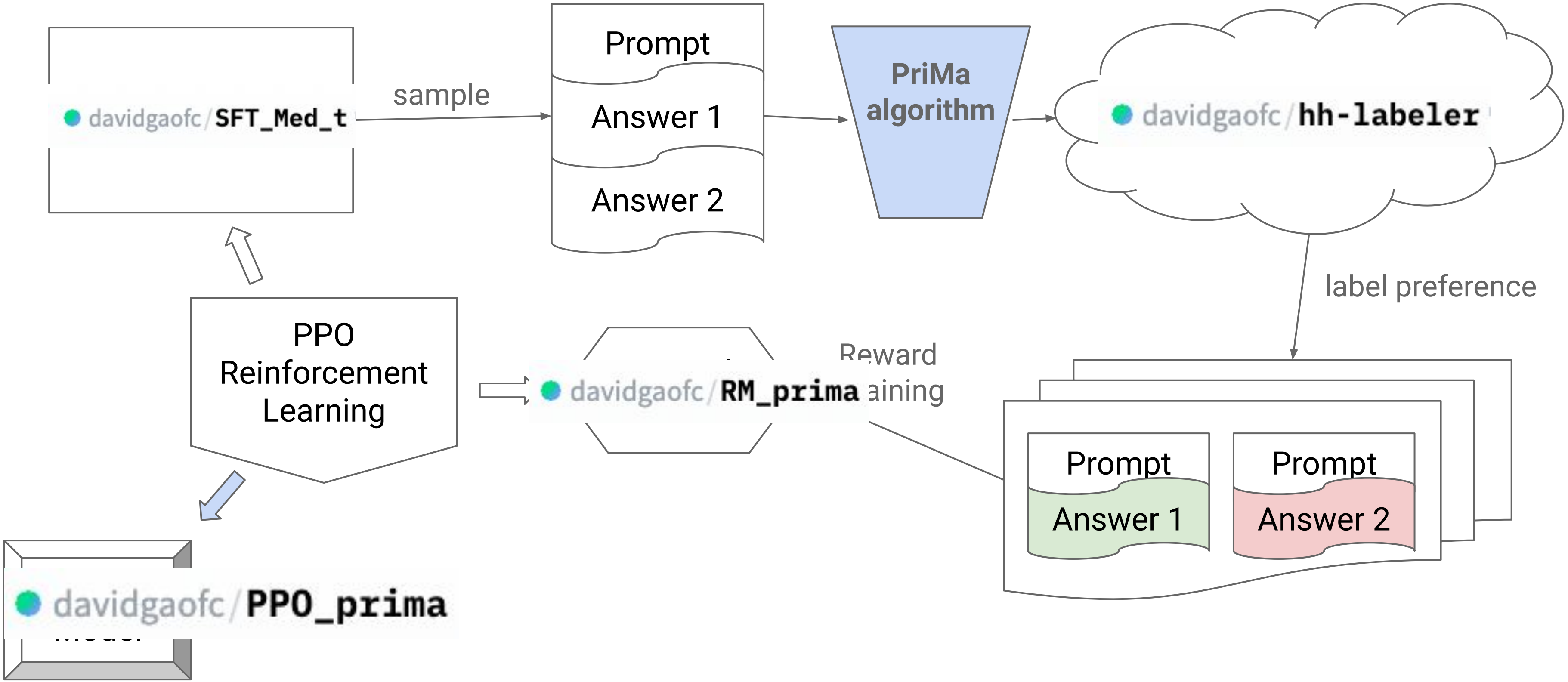


davidgaofc/PPO_prima

Privacy Preserving RLAIF

- Utilize the PPO Algorithm
- Use the Privacy Preserving Reward Model as RL signal
- Train using the same subsection of dataset designated for RL

Privacy Preserving RLAIF Pipeline



Attack Model

- Shadow Attack Overview
- Shadow Model
- Attack Model
- Attack Results

Shadow Attacks

- proposed by Shokri *et al.*
- Attacker trains a model as similar as they can to the original model
- Using the model's response to in-sample and out-of-sample data points, determine membership inference

davidgaofc/SFT_shadow

Shadow Model

- We fine-tune a second SFT model on the same architecture as the first, T5-Small
- Disjoint set of training data from MedQuAD dataset

Attack Model

- Using Shadow model's responses to in-sample and out-of-sample responses, we train classifier for membership inference
- DistilBERT

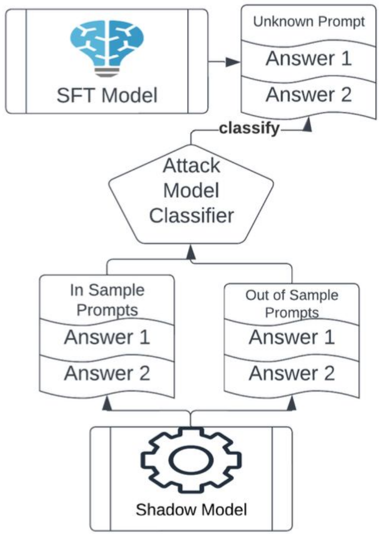


Figure 4. A diagram showing the Shadow attack for membership inference.

Attack Results

- Overall, precision is pretty low due to the complexity of the classification task
- Data run through PriMa (30%, 1 iteration) decreases the precision of attack

	Base Data	**PriMa Data**
Attack Precision	50.19	49.37

Table 1. Membership Inference Attack Results

Model Evaluations

- ROUGE scores
- Pairwise Alignment Win Rates

ROUGE

- ROUGE scores on the test dataset are mostly stable
- Decreases in ROUGE-1 and ROUGE-L are expected

	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.

Pairwise Alignment Comparisons

- Main goal of RLAIF - alignment
- Reuse AI Preference Labeler Model
- PriMa RLAIF actually aligns more than the Vanilla RLAIF

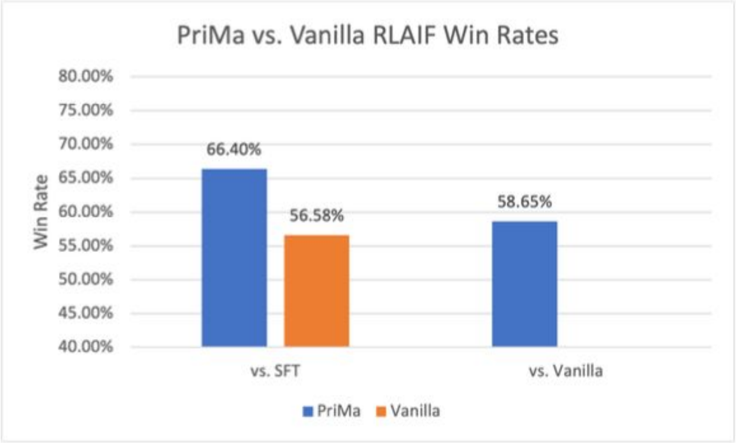


Figure 5. Results from head to head annotations by the AI Preference Labeler.

Discussion

- Novelty
- Limitations

Novelty

- No other studies found that explores privacy in RLAIF (since the concept is very new)
- Improvement against membership inference
- Alignment improvement
- Accuracy consistency

Limitations

- Dataset size
- Computational restrictions
- Transferability
  - LLMs rely on data (unsure if this is applicable in other contexts)

Demo

(if time permits)

Thanks!