Abstract

This thesis concerns the statistical evaluation of reward models used in language modeling. A reward model is a function that takes a prompt and a response and assigns a score indicating how 'good' that response is for the prompt. A key challenge is that reward models are usually imperfect proxies for actual preferences. For example, we may worry that a model trained to reward helpfulness learns to instead prefer longer responses. In this thesis, we develop an evaluation method, RATE (Rewrite-based Attribute Treatment Estimators), that allows us to measure the *causal* effect of a given attribute of a response (e.g., length) on the reward assigned to that response. The core idea is to use large language models (LLMs) to rewrite responses to produce imperfect counterfactuals, and to adjust for rewriting error by rewriting *twice*. We prove $\sqrt{n}$-consistency of the estimator under reasonable assumptions and demonstrate its effectiveness empirically. This work extends classical causal inference techniques to handle the unique challenges posed by modern language models.

This thesis is based on joint work with David Reber, Todd Nief, Cristina Garbacea, and Victor Veitch. The statistical theory, consistency proofs, and methodological extensions presented here represent my primary contributions to this collaboration.

1 Introduction

1.1 Motivation

In the context of large language models (LLMs), reward models evaluate the quality or appropriateness of model outputs, either by assessing individual responses or comparing multiple alternatives. Such models are useful in a variety of settings, including alignment of LLMs, ranking output samples (e.g., to use in a best-of-n sampling procedure), or evaluation of LLM performance.

Ideally, reward models would directly and perfectly measure whatever aspect of the output is important—e.g., we might have a reward for mathematical problem solving based on whether the generated response is correct. However, reward models are commonly learned from training data that imperfectly measures somewhat nebulous attributes. For example, a common task is to train a reward model based on human preferences for which of two responses is better. This results in a challenge where, even with a reward model in hand, we are not certain what it is actually rewarding. For example, we might worry that a model trained to reward helpfulness learns to instead simply prefer longer responses.

1.2 Statistical Challenges in Text Model Evaluation

Language models present unique statistical challenges for causal inference. The data consists of high-dimensional, unstructured text where multiple attributes may be correlated in complex ways. A straightforward approach would be to collect a dataset of prompt/response pairs, label each response as having or not having the attribute of interest, and then compare the average reward assigned to responses with and without the attribute. However, this approach has the limitation that it does not account for 'spurious' correlations that may exist in the data. For example, it may be that longer responses are more likely to be helpful (even though simply making a response longer does not necessarily make it more helpful). We call this correlation 'spurious' because helpfulness and length are not causally related (in the sense that neither is necessarily a cause of the other).

If we then applied the straightforward approach to this data to assess whether a given model is rewarding helpfulness, we would conclude that it is *even if the model only rewards length and is indifferent to helpfulness.* If we then used this reward model as a proxy for helpfulness in a downstream alignment task, then the actual effect of alignment would be to make responses longer, without (necessarily) affecting helpfulness.

1.3 Causal Inference Framework

Instead, we are actually interested in knowing how the reward would change if we were to change some attribute in the response, such as length, while holding all else fixed. This is the *causal* effect of the attribute on the reward. The core statistical challenge is that we only observe one potential outcome for each response - we cannot simultaneously observe how the reward model would score both a helpful and unhelpful version of the exact same response. This is the fundamental problem of causal inference.

Moreover, any attempt to modify a response to change one attribute (like helpfulness) may inadvertently affect other attributes as well. This requires careful consideration of:

- How to properly identify and estimate treatment effects when we only observe one potential outcome
- How to account for the fact that our treatment (changing an attribute) may affect multiple aspects of the text
- How to construct and leverage approximate counterfactuals

The contribution of this work is to develop and demonstrate a statistically principled method for estimating the causal effect of an attribute of a response on the reward assigned to that response. Our key methodological insight is using rewrites of rewrites to remove bias from imperfect counterfactuals.

2 Statistical Framework

2.1 Setup

Reward models are typically implemented in two ways:

1. As functions $R(x, y)$ that take a prompt x and a response y as inputs and return a real number indicating the quality of the response for the prompt.
2. As functions $R(x, y_1, y_0)$ that take a prompt x and two responses y_1 and y_0 as inputs and return a real number describing the relative quality of y_1 compared to y_0.

Our results apply to both implementations, but we focus on the first for clarity. Suppose we have a dataset of prompt-completion pairs $\{(x^i, y^{ij})\}$, where the x^i are prompts and the y^{ij} are completions (also referred to as 'responses'). We have a reward model $R(x^i, y^{ij})$ that assigns a scalar reward to each prompt-completion pair. We are interested in understanding how the reward model responds to a certain attribute, represented by the function W, within the completions. For each prompt-completion pair, we have a binary label $w^{ij} = W(x^i, y^{ij}) \in \{0, 1\}$ indicating whether the completion has the attribute of interest.

For example, W might represent helpfulness, which varies based on the context given by the prompt. A recipe could be helpful for questions about cooking but not for questions about history. We focus on binary attributes for simplicity—many attributes of interest (such as length) can often be naturally binarized.

2.2 Treatment Effects

To isolate the effect of a given attribute on the reward model, we take a causal perspective. Concretely, we can formalize the responsiveness of a reward model to some attribute W through several treatment effects:

Definition 2.1 (Average Treatment Effect)**.** The average treatment effect (ATE) of attribute W on reward R is:

$$\text{ATE} = \mathbb{E}[R(X, Y(1)) - R(X, Y(0))]$$

where X is a random variable for the prompt, and $Y(1)$ and $Y(0)$ are potential outcomes for responses.

This quantity represents the expected change in reward if we were to change the attribute W from 0 to 1, while keeping all other aspects of the response fixed. The pair of responses $(Y(0), Y(1))$ should be identical in all aspects except for the attribute W—e.g., if W is helpfulness then each counterfactual response should have the same writing level, sentiment, topic, etc.

Beyond the ATE, we also consider:

Definition 2.2 (Average Treatment Effect on the Treated and Untreated)**.** The average treatment effect on the treated (ATT) and untreated (ATU) are defined as:

$$\text{ATT} = \mathbb{E}[R(X, Y(1)) - R(X, Y(0))|W = 1]$$
$$\text{ATU} = \mathbb{E}[R(X, Y(1)) - R(X, Y(0))|W = 0]$$

These estimands capture asymmetric effects: if $W = 1$ is a helpful response, the ATT measures the change in reward when we take a helpful response and make it less helpful, and the ATU measures the change in reward when we take an unhelpful response and make it more helpful. These estimands can differ substantially from each other, so some thought should be given to which is most relevant to the question at hand. Indeed, even human preferences are often asymmetric, so we might expect reward model preferences to be as well.

2.3 Challenges in Estimation

A naive approach to measuring the sensitivity of a reward model to an attribute would be to take the dataset of prompt-completion pairs, label each completion as having or not having the attribute, then check whether the rewards differ between the groups. Mathematically, we might estimate:

$$\hat{\tau}_{\text{naive}} = \frac{1}{n_1} \sum_{(x^i, y^{ij}): w^{ij}=1} R(x^i, y^{ij}) - \frac{1}{n_0} \sum_{(x^i, y^{ij}): w^{ik}=0} R(x^i, y^{ik})$$

where n_1 and n_0 are the numbers of examples with $W = 1$ and $W = 0$, respectively.

This can be viewed as a finite sample estimator for:

$$\mathbb{E}[R(X, Y)|W = 1] - \mathbb{E}[R(X, Y)|W = 0]$$

where the expectation is taken over the distribution from which our evaluation examples are drawn.

However, even in the infinite data limit, this quantity does not generally isolate the effect of W on R. For instance, if helpful responses tend to be longer, then the naive estimator will conflate the effects of helpfulness and length.

A common solution in causal inference is to use backdoor adjustment, conditioning on confounding variables to isolate the causal effect. However, this approach faces significant challenges when working with text:

- Text data is extremely high-dimensional, making it difficult to identify and condition on all relevant confounders
- Many text attributes are nebulous and hard to quantify (e.g., writing quality, coherence)
- The relationship between attributes can be complex and context-dependent, making it difficult to identify which attributes are confounders

These challenges motivate our development of RATE, which uses LLMs to directly generate approximate counterfactuals rather than attempting to adjust for confounders. The key insight is that we can use rewrites of rewrites to correct for imperfections in these generated counterfactuals.

3 RATE: Rewrite-based Attribute Treatment Estimators

3.1 Core Approach

Whatever our choice of estimand, we need a method to estimate it. Here, we develop a method, RATE, that uses rewrites to estimate the causal effect of an attribute on a reward model. The core idea is to create pairs of responses where the only difference is in the attribute of interest. For example, we might modify a response to change its sentiment from positive to negative, while keeping all other aspects of the response the same. The goal is for these modified responses to directly approximate the unobserved counterfactual responses.

3.1.1 Rewrites with LLMs

In practice, we implement rewrites using an LLM. We begin with a labeled dataset containing ground truth binary variables for attributes such as complexity, sentiment, or helpfulness. We then instruct the LLM to rewrite the responses to the opposite state of the binary variable. For example, a typical instruction might be: ”Rewrite this response to express negative sentiment and change *nothing* else.”

We use $\mathrm{Re}(x^i, y^{ij}, w)$ to denote the rewrite operation, which takes a prompt-response pair (x^i, y^{ij}) and a desired attribute value w, returning a modified response $\tilde{y}^{ij}$ such that $W(x^i, \tilde{y}^{ij}) = w$.

3.1.2 Rewrite Instructions

There is significant flexibility in how to instruct an LLM to rewrite. For instance, when rewriting for ’helpfulness’, we might instruct the LLM to ”Rewrite this response to be more helpful”, or instruct it to ”Rewrite this response to be more helpful, providing additional relevant information or clarification.” In this example, the second instruction makes the meaning of ’helpful’ more precise. Generally, changing the instruction changes the nature of the rewrites generated, and thus changes the attribute that is being modified.

This is inevitable. Ambiguity in interventions is unavoidable in causal inference. In our context, there is subjectivity in what helpfulness, complexity, or sentiment actually mean. An advantage of the rewrite approach is that it allows us to use natural language to specify, as clearly as possible, what property we are actually trying to modify. We can understand whether our instructions are having the intended effect by qualitatively examining the rewritten outputs and checking that they vary the attribute of interest while leaving the rest of the response unchanged.

3.2 Imperfect Rewrites

If the rewrites produced perfect counterfactuals, it would then be straightforward to estimate the causal effect of the attributes. Namely, we could compare the rewards of the original responses to the rewards of the rewrites. However, the rewrites are often imperfect, modifying off-target attributes. These off-target modifications may affect the reward, causing the simple comparison to be misleading.

To analyze this formally, we introduce a latent variable model:

Definition 3.1 (Latent Variable Model)**.** A response Y is modeled as:

$$Y = Y(W, Z, \xi)$$

where:

- Y is the observed response
- W is the target attribute we aim to manipulate (e.g., sentiment)
- Z represents off-target attributes that are invariant to rewrites (e.g., topic)
- ξ represents off-target attributes that may be affected by rewrites (e.g. grammar)

This decomposition allows us to reason about rewrite errors. Whenever we rewrite some response y^{ij} (to $W = w$), we introduce some error ϵ_w^{ij} in the reward because of our inability to perfectly produce the counterfactual $y^{ij}(w)$:

$$\epsilon_w^{ij} = R(x^i, \mathrm{Re}(x^i, y^{ij}, w)) - R(x^i, y^{ij}(w))$$

We would like to correct for these errors. Yet the whole point of the rewrites is to approximate the counterfactuals $y^{ij}(w)$, so we cannot directly measure ϵ_w^{ij}.

3.3 RATE Procedure

Surprisingly, the solution is to introduce *more noise.* Instead of comparing a rewrite to the original response, we compare it to the rewrite of the rewrite, thereby canceling out off-target noise introduced by the rewrite process. That is, rather than selecting (original, rewrite):

$$\tilde{\tau}^{ij} = \begin{cases} R(x^i, y^{ij}) - R(x^i, \mathrm{Re}(x^i, y^{ij}, 0)), & \text{if } w^{ij} = 1 \\ R(x^i, \mathrm{Re}(x^i, y^{ij}, 1)) - R(x^i, y^{ij}), & \text{if } w^{ij} = 0 \end{cases}$$

we instead compare the (rewrites, rewrites of rewrites) pairs:

$$\hat{\tau}^{ij} = \begin{cases} R(x^i, \mathrm{Re}(\mathrm{Re}(x^i, y^{ij}, 0), 1)) - R(x^i, \mathrm{Re}(x^i, y^{ij}, 0)), & \text{if } w^{ij} = 1 \\ R(x^i, \mathrm{Re}(x^i, y^{ij}, 1)) - R(x^i, \mathrm{Re}(\mathrm{Re}(x^i, y^{ij}, 1), 0)), & \text{if } w^{ij} = 0 \end{cases}$$

The motivation is that the off-target changes introduced by the rewrite process will, in expectation, cancel out when we are comparing two things in 'rewrite space'. For example, the tendency for LLMs to produce well-formatted text will affect both the first rewrite and the rewrite of the rewrite, so the overall contribution of this off-target change will cancel out. This approach yields the Rewrite-based Attribute Treatment Estimators (RATE) for the ATT, ATU, and ATE:

Algorithm 1 RATE: Rewrite-based Attribute Treatment Estimators

```
**Input:** Dataset $\{(x^i, y^{ij}, w^{ij})\}$, reward model $R$, function $\mathrm{Re}()$
**Return:** Estimates $\hat{\tau}_{\mathrm{ATT}}$, $\hat{\tau}_{\mathrm{ATU}}$, $\hat{\tau}_{\mathrm{ATE}}$
Initialize $n_1 \leftarrow \sum_{i,j} \mathbb{I}[w^{ij} = 1]$, $n_0 \leftarrow \sum_{i,j} \mathbb{I}[w^{ij} = 0]$
$\hat{\tau}_{\mathrm{ATT}} \leftarrow \frac{1}{n_1} \sum_{(i,j): w^{ij}=1} [R(x^i, \mathrm{Re}(\mathrm{Re}(x^i, y^{ij}, 0), 1)) - R(x^i, \mathrm{Re}(x^i, y^{ij}, 0))]$
$\hat{\tau}_{\mathrm{ATU}} \leftarrow \frac{1}{n_0} \sum_{(i,j): w^{ij}=0} [R(x^i, \mathrm{Re}(x^i, y^{ij}, 1)) - R(x^i, \mathrm{Re}(\mathrm{Re}(x^i, y^{ij}, 1), 0))]$
$\hat{\tau}_{\mathrm{ATE}} \leftarrow \frac{n_1}{n_0+n_1} \hat{\tau}_{\mathrm{ATT}} + \frac{n_0}{n_0+n_1} \hat{\tau}_{\mathrm{ATU}}$
**return** $\hat{\tau}_{\mathrm{ATT}}$, $\hat{\tau}_{\mathrm{ATU}}$, $\hat{\tau}_{\mathrm{ATE}}$
```

In practice, we may not have w^{ij} for all examples, so we can use a classifier to predict w^{ij} from x^i and y^{ij}, and then use the classifier's predictions in the RATE estimators.

4 Theoretical Analysis of RATE

Under reasonable assumptions, RATE is a $\sqrt{n}$-consistent estimator of the average treatment effect.

Latent Variable Model To analyze the rewrite operation, we need to conceptualize how different aspects of a response might change during rewriting. Imagine a response as having three types of attributes: the target attribute we want to change (like sentiment), attributes that should remain constant (like topic), and attributes that might unintentionally change (like specific wording). We can formalize this idea using a latent variable model:

$$Y = Y(W, Z, \xi)$$

where:

- Y is the observed response
- W is the target attribute we aim to manipulate (e.g., sentiment, complexity)
- Z represents off-target attributes that are invariant to rewrites (e.g., topic, language)
- ξ represents off-target attributes that may be affected by rewrites (e.g. grammatical structure)

Intuitively, we expect some off-target attributes Z to remain unchanged during rewrites. For example, if we ask an LLM to change the sentiment of an English text, we don't expect it to suddenly produce Korean. However, other off-target attributes ξ may change: for instance, grammar and punctuation might be corrected.

Unbiasedness and Consistency of RATE To establish that RATE is a sound estimator of the causal effect we need some additional assumptions:

1. We assume that the reward model can be decomposed additively:

   $$R(X, Y(W, Z, \xi)) = R_{W,Z}(X, W, Z) + R_{\xi}(X, \xi)$$

   where:

   (a) $R_{W,Z}(X, W, Z)$ is the component of the reward that depends on the target attribute W and the immutable off-target attributes Z.

   (b) $R_{\xi}(X, \xi)$ is the component of the reward that depends on the mutable off-target attributes ξ.

   This means that we don't need to worry about potential interactions between rewrite errors (affecting ξ) and other attributes of the response (Z), even if W and Z have interactions. Some justification for this assumption is that, intuitively, human preferences for many attributes are separable. For example, the strength of our preference for a response to be helpful (W) is unlikely to depend on attributes like the specific wording used (ξ). Rewards, then, as approximations of human preferences, should also be separable in this way. To be sure, such separability does not, intuitively, hold in some cases (e.g., the strength of our preference for a response to be cheerful may depend on the topic of the response), but these cases seem to involve immutable attributes Z rather than mutable attributes ξ, at least when we are considering rewrites done by sophisticated LLMs, as they will not change the topic of a response when asked to change its sentiment.

2. We assume that the off-target changes introduced by the rewrite process are randomly drawn from a distribution determined by the particular rewrite method being used. That is,

$$\mathrm{Re}(Y(W,Z,\xi)) \overset{d}{=} Y(W,Z,\tilde{\xi}), \quad \text{where } \tilde{\xi} \overset{d}{\sim} P_{\mathrm{Re}}(\tilde{\xi})$$

   For example, when our rewriter is GPT-4o, the off-target yet mutable attributes such as specific word choice and grammatical structure are drawn from a 'GPT-like' distribution.

These assumptions lead to the following result:

Theorem 4.1 (Unbiasedness and Consistency)**.** *Let* $R(X, Y(W,Z,\xi)) = R_{W,Z}(X,W,Z) + R_{\xi}(X,\xi)$ *and* $Re(Y(W,Z,\xi)) \overset{d}{=} Y(W,Z,\tilde{\xi})$ *where* $\tilde{\xi} \overset{d}{\sim} P_{Re}(\tilde{\xi})$*. Assume that the reward function is bounded. Then the RATE estimators, defined as:*

$$\begin{aligned}
\hat{\tau}_{ATT} &= \frac{1}{n_1} \sum_{(i,j):w^{ij}=1} [R(x^i, Re(Re(x^i, y^{ij}, 0), 1)) - R(x^i, Re(x^i, y^{ij}, 0))] \\
\hat{\tau}_{ATU} &= \frac{1}{n_0} \sum_{(i,j):w^{ij}=0} [R(x^i, Re(x^i, y^{ij}, 1)) - R(x^i, Re(Re(x^i, y^{ij}, 1), 0))] \\
\hat{\tau}_{ATE} &= \frac{n_1}{n_0+n_1}\hat{\tau}_{ATT} + \frac{n_0}{n_0+n_1}\hat{\tau}_{ATU}
\end{aligned}$$

where n_1 *and* n_0 *are the number of pairs with observed* $W=1$ *and* $W=0$ *respectively, are unbiased and* $\sqrt{n}$*-consistent estimators of the ATT, ATU, and ATE.*

See Appendix A for the proof.

5 Empirical Evaluation

In our paper, we evaluate popular reward models using RATE on real-world and synthetic data. Experiments show:

- Across a variety of attributes and datasets, RATE gives substantively different estimates compared to the naive (non-causal) baseline.
- In semi-synthetic data with known ground truth behavior, RATE is robust to distributional shift, while the naive estimator is not.
- Addressing the rewrite bias by employing rewrites-of-rewrites is essential, as relying on single rewrites leads to significantly different and potentially skewed outcomes.

As the experiments are not the primary focus of this thesis, we refer the reader to the full paper for more details.

6 Discussion

Generalization to Contrastive Rewards The RATE procedure applies more generally to contrastive rewards of the form $R(x, y_1, y_0)$, which assign a relative reward for y_1 compared to y_0. In this case, RATE enables us to compute $\mathbb{E}[R(X, Y(W=1), Y(W=0))]$, the expected increase in relative reward attributable to changing attribute W in isolation of everything else, simply by replacing the summands in the earlier formulations. Specifically, we can modify our RATE estimators as follows:

$$\begin{aligned}
\hat{\tau}_{\text{ATT}} &= \frac{1}{n_1} \sum_{(i,j):w^{ij}=1} [R(x^i, \text{Re}(\text{Re}(x^i, y^{ij}, 0), 1), \text{Re}(x^i, y^{ij}, 0))] \\
\hat{\tau}_{\text{ATU}} &= \frac{1}{n_0} \sum_{(i,j):w^{ij}=0} [R(x^i, \text{Re}(x^i, y^{ij}, 1), \text{Re}(\text{Re}(x^i, y^{ij}, 1), 0))] \\
\hat{\tau}_{\text{ATE}} &= \frac{n_1}{n_0 + n_1} \hat{\tau}_{\text{ATT}} + \frac{n_0}{n_0 + n_1} \hat{\tau}_{\text{ATU}}
\end{aligned}$$

As an example of why this may be useful, consider an evaluator LLM that takes a prompt and two responses and returns a preference for which is better. We may view this as a contrastive reward with outputs in $0, 1$. RATE enables us to estimate how sensitive this evaluator is to different attributes considered in isolation. Notice that in the particular case where $R(x, y_1, y_0) = R(x, y_1) - R(x, y_0)$, as in, e.g., the Bradley-Terry model, the contrastive RATE estimate is the same as the pointwise RATE estimate described in the main body of the paper. This highlights the versatility of our approach, as it naturally extends to both pointwise and contrastive reward models.

Generalization to Model Edits Note that we can construct a "reward function" for a model edit by comparing the original model to the edited model. In particular, we could define

$$\tilde{R}_{\pi,\pi_0}(x, y) = \log \frac{\pi(y|x)}{\pi_0(y|x)}$$

where $\pi(y|x)$ is the probability of generating y given x under the edited model, and $\pi_0(y|x)$ is the probability under the original model. For instance, we could determine whether we have successfully fine-tuned a model to be more friendly by estimating the ATE of friendliness on the log-likelihood ratio relative to the baseline model,

$$\text{ATE}_{W,\pi,\pi_0} = \mathbb{E}[\tilde{R}_{\pi,\pi_0}(X, Y(W=1)) - \tilde{R}_{\pi,\pi_0}(X, Y(W=0))]$$

where $Y(W=1)$ and $Y(W=0)$ are counterfactuals differing only in friendliness.

Notice that this is different from the naive approach of comparing the outputs of the original and edited models to see which is more friendly, as this may be confounded by the fine-tuned model's drift on other attributes correlated with friendliness. Instead, the causal framing allows us to isolate a single attribute and determine whether the model has been successfully fine-tuned on this dimension. Just as we have done with real-world reward models, the RATE method allows us to estimate this ATE by rewriting the responses to change only the friendliness attribute, and then comparing the log-likelihood ratios of the original and rewritten responses under the original and edited models.

This could be particularly useful in the context of model interpretability. For instance, if we believe that a vector λ is a "steering vector" for friendliness and define π_λ as probability distribution over tokens induced by adding λ to the residual stream, we could see whether the ATE with respect to friendliness is non-zero, $\text{ATE}_{\pi_\lambda,\pi_0,W} \neq 0$. This would suggest that λ is indeed steering the model towards friendliness. For a targeted steering vector, we would like the ATE with respect to all other attributes to be zero, as the steering vector is only intended to affect friendliness.

A Proofs

Theorem 4.1 (Unbiasedness and Consistency)**.** *Let $R(X, Y(W, Z, \xi)) = R_{W,Z}(X, W, Z) + R_{\xi}(X, \xi)$ and $Re(Y(W, Z, \xi)) \overset{d}{=} Y(W, Z, \tilde{\xi})$ where $\tilde{\xi} \overset{d}{\sim} P_{Re}(\tilde{\xi})$. Assume that the reward function is bounded. Then the RATE estimators, defined as:*

$$
\begin{aligned}
\hat{\tau}_{ATT} &= \frac{1}{n_1} \sum_{(i,j):w^{ij}=1} [R(x^i, Re(Re(x^i, y^{ij}, 0), 1)) - R(x^i, Re(x^i, y^{ij}, 0))] \\
\hat{\tau}_{ATU} &= \frac{1}{n_0} \sum_{(i,j):w^{ij}=0} [R(x^i, Re(x^i, y^{ij}, 1)) - R(x^i, Re(Re(x^i, y^{ij}, 1), 0))] \\
\hat{\tau}_{ATE} &= \frac{n_1}{n_0 + n_1} \hat{\tau}_{ATT} + \frac{n_0}{n_0 + n_1} \hat{\tau}_{ATU}
\end{aligned}
$$

where n_1 and n_0 are the number of pairs with observed $W = 1$ and $W = 0$ respectively, are unbiased and $\sqrt{n}$-consistent estimators of the ATT, ATU, and ATE.

Proof. First, we'll prove unbiasedness and $\sqrt{n_1}$-consistency of $\hat{\tau}_{\text{ATT}}$ and $\sqrt{n_0}$-consistency of $\hat{\tau}_{\text{ATU}}$, and then use these results for $\hat{\tau}_{\text{ATE}}$. Throughout, we use $\tilde{\xi}$ and $\tilde{\tilde{\xi}}$ to denote i.i.d. samples from the distribution P_{ξ}, where the former comes from the first rewrite and the latter from the rewrite of the rewrite.

1. Unbiasedness and Consistency of $\hat{\tau}_{\mathbf{ATT}}$

Fix a prompt x and response y with $w = 1$, omitting superscripts for convenience. We calculate:

$$
R(x, \text{Re}(\text{Re}(x, y, 0), 1)) - R(x, \text{Re}(x, y, 0))
$$

which has expected value:

$$
\begin{aligned}
\mathbb{E}[R(x, \text{Re}(\text{Re}(x, y, 0), 1)) - R(x, \text{Re}(x, y, 0))] &= \mathbb{E}_{\tilde{\xi}, \tilde{\tilde{\xi}} \sim P_{\text{Re}}}[R(x, y(1, z, \tilde{\tilde{\xi}})) - R(x, y(0, z, \tilde{\xi}))] \\
&= \mathbb{E}_{\tilde{\xi}, \tilde{\tilde{\xi}} \sim P_{\text{Re}}}[R_{W,Z}(x, 1, z) + R_{\xi}(x, \tilde{\tilde{\xi}}) - R_{W,Z}(x, 0, z) - R_{\xi}(x, \tilde{\xi})] \\
&= R_{W,Z}(x, 1, z) - R_{W,Z}(x, 0, z) \\
&= R_{W,Z}(x, 1, z) - R_{W,Z}(x, 0, z) + R_{\xi}(x, \xi) - R_{\xi}(x, \xi) \\
&= R(x, y(1, z, \xi)) - R(x, y(0, z, \xi)) \\
&= R(x, y(1)) - R(x, y(0))
\end{aligned}
$$

Therefore, as an average over these quantities, we have:

$$
\mathbb{E}[\hat{\tau}_{\text{ATT}}] = \mathbb{E}[R(X, Y(1)) - R(X, Y(0)) | W = 1] = \text{ATT}
$$

For $\sqrt{n_1}$-consistency, by boundedness of R, define $\delta_i = R(x^i, \text{Re}(\text{Re}(x^i, y^{ij}, 0), 1)) - R(x^i, \text{Re}(x^i, y^{ij}, 0))$. Then for all i, $\text{Var}(\delta_i) \leq B$ for some constant B. Therefore:

$$
\text{Var}(\hat{\tau}_{\text{ATT}}) = \frac{1}{n_1^2} \sum_{(i,j):w^{ij}=1} \text{Var}(\delta_i) \leq \frac{B}{n_1}
$$

By Chebyshev's inequality:

$$
P(|\sqrt{n_1}(\hat{\tau}_{\text{ATT}} - \text{ATT})| > \epsilon) \leq \frac{B}{\epsilon^2} = O(1)
$$

Therefore $\sqrt{n_1}(\hat{\tau}_{\text{ATT}} - \text{ATT}) = O_p(1)$.

2. Unbiasedness and Consistency of $\hat{\tau}_{\mathbf{ATU}}$

The proof follows identically with $W = 0$, yielding $\sqrt{n_0}(\hat{\tau}_{\text{ATU}} - \text{ATU}) = O_p(1)$.

3. Unbiasedness and Consistency of $\hat{\tau}_{\mathbf{ATE}}$

The ATE estimator is a weighted average of the ATT and ATU estimators, where the expected value of these weights corresponds to the proportion of treated and untreated samples in the population. Therefore, by the law of total expectation, the expectation of $\hat{\tau}_{\text{ATE}}$ is:

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_{\text{ATE}}] &= \mathbb{E}[R(X, Y(1)) - R(X, Y(0))|W = 1] \cdot P(W = 1) \\
&\quad + \mathbb{E}[R(X, Y(1)) - R(X, Y(0))|W = 0] \cdot P(W = 0) \\
&= \mathbb{E}[R(X, Y(1)) - R(X, Y(0))] \\
&= \text{ATE}
\end{aligned}$$

Thus, $\hat{\tau}_{\text{ATE}}$ is an unbiased estimator of the ATE.

For $\sqrt{n}$-consistency of ATE, recall that:

$$\hat{\tau}_{\text{ATE}} = \frac{n_1}{n}\hat{\tau}_{\text{ATT}} + \frac{n_0}{n}\hat{\tau}_{\text{ATU}}$$

$$\text{ATE} = P(W = 1)\text{ATT} + P(W = 0)\text{ATU}$$

Since $\frac{n_1}{n} \to_p P(W = 1)$, $\frac{n_0}{n} \to_p P(W = 0)$, $\sqrt{n_1}(\hat{\tau}_{\text{ATT}} - \text{ATT}) = O_p(1)$, and $\sqrt{n_0}(\hat{\tau}_{\text{ATU}} - \text{ATU}) = O_p(1)$, by Slutsky's theorem:

$$\sqrt{n}(\hat{\tau}_{\text{ATE}} - \text{ATE}) = O_p(1)$$

□