# Degree Paper for Masters in Statistics

## — A Sample Format

Sean Richardson

Advisor: Victor Veitch

Approved _____

Date _____

January 5, 2025

**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 approximate 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 text data in the context of reward modeling.

*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.*

# Contents

# 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 evaluating LLM performance in a stable, automated way.

Ideally, reward models would perfectly capture and measure whatever aspect of the output is important—for example, a reward model for mathematical problem solving would precisely identify whether the generated response is correct. However, reward models are commonly learned from training data that imperfectly measures more nebulous attributes. For instance, a frequent approach is to train a reward model based on human pairwise preferences for which of two responses is "more helpful." This may lead the reward model to conflate *helpfulness* with other textual attributes it happens to correlate with (e.g., length, style, politeness).

**The Key Challenge.** Even if we have a reward model, it remains unclear what the model is *actually* rewarding. As a result, we risk deploying or further training a model on a misaligned or spurious proxy. For instance, a "helpfulness" reward model could inadvertently reward longer responses, regardless of their true helpfulness.

## 1.2 Statistical Challenges in Text-Model Evaluation

Language models introduce unique statistical challenges for causal inference:

1. **Spurious Correlations:** Any text attribute of interest (sentiment, politeness, helpfulness, complexity) might be confounded with other attributes in the data, making purely observational methods insufficient.

2. **High Dimensionality and Complexity:** Text is extremely high-dimensional and unstructured. Multiple textual attributes may be correlated in ways that are difficult to disentangle.

**A Naive Approach.** One might try to isolate the impact of an attribute $W$ by simply partitioning the data into $W = 1$ and $W = 0$ and comparing mean reward. However, if $W$ is correlated with other features of the text, that naive approach will conflate multiple causal effects, failing to isolate the *direct* effect of $W$.

## 1.3 Our Contribution

To properly estimate how the reward would change if we were to *intervene* on one specific text attribute while keeping the other text aspects fixed, we adopt a formal causal perspective. This thesis introduces **RATE (Rewrite-based Attribute Treatment Estimators)**, a novel approach that uses large language models (LLMs) to rewrite text into approximate counterfactuals. These counterfactuals differ only in one target attribute (e.g., length, sentiment, helpfulness), leaving other attributes as intact as possible.

However, LLM-based rewrites are typically *imperfect*, and may inadvertently alter additional attributes. To correct for these off-target changes, we propose rewriting *twice* and comparing the resulting reward differences. Under mild assumptions, this procedure yields *unbiased,* $\sqrt{n}$-consistent estimates of the average treatment effect (ATE), average treatment effect on the treated (ATT), and average treatment effect on the untreated (ATU).

### 1.4 Structure of this Thesis

In Section 2, we situate RATE among other causal approaches and reward-model evaluation methods. In Section 3, we formalize the problem setup, introducing notation and the concept of causal estimands in the text domain. Section 4 describes our main algorithm, RATE, and Section 5 presents theoretical justifications. Section 6 outlines empirical studies, highlighting both real-world and synthetic evaluations. Finally, Section 7 covers extensions such as contrastive rewards and model edits, and Section 8 presents broader conclusions and future directions.

## 2 Related Work

There is growing interest in understanding and mitigating biases or spurious correlations in NLP tasks, including reward modeling:

- **Causal Inference in NLP.** Researchers have explored causal-inference frameworks for text, especially for text classification [**?**]. *RATE* extends these ideas to *reward models*, which have a similar structure but differ in the sense that they are often used for reinforcement learning or ranking scenarios.
- **Counterfactual Generation.** Generating minimal textual edits to isolate one attribute is an active research area. Systems like Polyjuice [**?**] create diverse counterfactuals, but often do not explicitly correct for rewriting artifacts. CEBaB [**?**] manually constructs counterfactuals to isolate sentiment or other aspects, though it requires significant human effort. RATE automatically generates and *corrects* rewrites for spurious changes.
- **Reward Modeling and Evaluation.** [**?**] propose large-scale benchmarks for reward models without taking a causal perspective. [**?**] addresses the presence of certain biases (e.g., length) in reward models but does not use causal methods to estimate such biases. RATE aims to *quantify* and isolate these biases by a principled causal procedure.

Overall, RATE can be viewed as an extension of text-oriented causal inference methods to address the unique demands of reward modeling—specifically, we focus on rewriting text in a way that yields an *unbiased* measure of how changing an attribute will shift the reward.

# 3 Problem Setup and Causal Estimands

## 3.1 Reward Models

Reward models are typically implemented in two ways:

1. **Pointwise:** $R(x, y)$ takes a prompt $x$ and a response $y$ and returns a scalar reward.
2. **Contrastive:** $R(x, y_1, y_0)$ takes a prompt $x$ and two responses $y_1, y_0$, returning a scalar indicating which response is better (or by how much).

Our derivations hold for both, but we focus on the pointwise version to keep notation simple. Though the details of how a reward model is implemented are not essential, some background context may be helpful. Typically, an LLM is presented with a prompt $x$ and produces for each prompt a pair of responses $y_1, y_0$. Human raters are then asked to compare the two responses. Without loss of generality we can call $y_1$ the preferred response, and $y_0$ the non-preferred response. The Bradley-Terry model is a common choice for $\Pr(y_1 > y_0)$ where $>$ is the preference relation:

$$\Pr(y_1 > y_0) = \sigma(R^*(x, y_1) - R^*(x, y_0)) \tag{1}$$

where $\sigma$ is the logistic function and $R^*(x, y)$ is the oracle reward function (i.e., some underlying function explaining the human preference). A parametric reward model $R_\theta(x, y)$ is then trained to predict $\Pr(y_1 > y_0)$ by maximum likelihood estimation:

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^{n} \log \sigma(R_\theta(x^i, y_1^i) - R_\theta(x^i, y_0^i)) \tag{2}$$

## 3.2 Notation

Let $\{(x^i, y^i)\}$ be a dataset of prompt-response pairs, with $R(x^i, y^i) \in \mathbb{R}$.

We consider a binary attribute $W(x^i, y^i) \in \{0, 1\}$ (e.g., "sentiment is positive" or "response is long") that we hypothesize may affect the reward. For instance, in a user-query dataset, $W$ might indicate whether the response is helpful for that query.

## 3.3 Naive Estimation and Its Pitfall

An obvious approach to see if $W$ influences $R$ is to compare the average reward for $W = 1$ vs. $W = 0$:

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

However, this fails to isolate the causal effect if $W$ is correlated with other confounding attributes. For example, if $W = 1$ responses are systematically longer or more detailed in ways that also raise their reward, we cannot infer whether the reward model truly favors $W$ (versus favoring those other correlated attributes).

### 3.4 Treatment Effects

To disentangle such confounders, we adopt a causal-inference framework. For each $(x, y)$, suppose there exist two *potential responses*:

$$Y(0), \quad Y(1),$$

which are identical except that one has $W = 0$ and the other has $W = 1$. The *average treatment effect* (ATE) is:

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

We also define:

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

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

In principle, $Y(1)$ and $Y(0)$ are never simultaneously observed. We only see whichever label $W$ the data happened to have. This is the fundamental problem of causal inference [**?**].

### 3.5 Why Not Simple Backdoor Adjustments?

A common approach to causal inference is to apply *backdoor adjustments* (e.g., regression, matching, inverse propensity weighting) by conditioning on all confounders that causally affect both the treatment and the outcome [**?**]. However, in high-dimensional text settings, this approach faces critical barriers:

- **Unknown Causal Structure.** Language is unstructured and can contain innumerable latent factors (sentiment, style, topic, complexity, speaker identity, etc.) whose causal roles are often unclear. Determining which textual features truly act as confounders is therefore nontrivial, if not infeasible, without extensive domain knowledge or perfect annotation of these features.
- **Context-Dependence.** Even if we attempt to identify potential confounders (e.g., style, topic), these may vary across documents and domains. A feature that is a confounder in one context (e.g., sentiment in a product review) may be irrelevant in another (e.g., a factual Q&A). Consequently, there is no universal set of textual covariates we can reliably adjust for.
- **Combinatorial Explosion.** In principle, we could embed each text into a high-dimensional representation (e.g., word embeddings or bag-of-words) and treat each dimension as a covariate for regression or propensity modeling. However, this quickly becomes computationally intractable and statistically fragile ("curse of dimensionality"), especially when sample sizes are finite. This also introduces overlap issues, where the covariates may have different supports across treatment groups.

In contrast, a *rewrite-based* approach such as **RATE** circumvents the need to explicitly identify and condition on all relevant confounders. Instead, it leverages the power of LLMs to generate counterfactuals that are as close as possible to the desired treatment, while keeping other aspects of the text fixed.

# 4 RATE: Rewrite-based Attribute Treatment Estimators

## 4.1 Rewrite Operations with LLMs

To approximate the counterfactual $Y(1)$ for a text that actually had $W = 0$, we ask an LLM to *rewrite* $y$ to have $W = 1$ while leaving everything else the same. Denote:

$$\mathrm{Re}(x^i, y^i, w_{\text{new}}) \quad \longrightarrow \quad \tilde{y}^i,$$

where $\tilde{y}^i$ is the LLM-generated rewrite of $y^i$ that attempts to enforce $W(\tilde{y}^i) = w_{\text{new}}$ but otherwise preserve the rest of the text. An example instruction might be:

> "Rewrite the following text to be more helpful, *without* changing other aspects such as tone, sentiment, or factual content."

## 4.2 Imperfect Rewrites and the "Double Rewrite" Idea

**Challenge:** LLM rewrites can inadvertently alter off-target attributes (e.g., grammar, length, writing style). Let $\epsilon_w$ be the difference in the reward caused by these unintended changes. If we compare

$$R(x, y) - R\big(x, \mathrm{Re}(x, y, 0)\big),$$

we pick up the effect of rewriting itself, not purely the effect of the attribute.

**Solution:** Use *Rewrite-of-Rewrite*. Instead of comparing an original $y$ to one rewrite, we:

- Start with $y$ that has $W = 1$,
- Rewrite it to $W = 0$, then
- Rewrite that rewrite back to $W = 1$.

Symbolically,

$$\mathrm{Re}\big(x, \mathrm{Re}(x, y, 0), 1\big),$$

is the *rewrite-of-rewrite* from $W = 0$ back to $W = 1$. Intuitively, the textual artifacts introduced by rewriting once are also introduced by rewriting again, so they can cancel out in expectation. Hence the **RATE** estimator compares:

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

when the *original* $y$ has $W = 1$. Similarly, if $W = 0$, we compare $R(\mathrm{Re}(x, y, 1))$ to $R(\mathrm{Re}(x, \mathrm{Re}(x, y, 1), 0))$. We can define the RATE estimation procedure as follows:

**Algorithm 1** RATE: Rewrite-based Attribute Treatment Estimators

**Input:** Dataset $\{(x^i, y^i, w^i)\}$, reward model $R$, rewrite function Re

**Return:** Estimates $\widehat{\mathrm{ATT}}_{\mathrm{RATE}}$, $\widehat{\mathrm{ATU}}_{\mathrm{RATE}}$, and $\widehat{\mathrm{ATE}}_{\mathrm{RATE}}$

**Compute:** $n_1 = \sum_i \mathbf{1}[w^i = 1], \quad n_0 = \sum_i \mathbf{1}[w^i = 0]$

**For all** i: **obtain rewrites** $\mathrm{Re}(x^i, y^i, 0)$ if $w^i = 1$ and $\mathrm{Re}(x^i, y^i, 1)$ if $w^i = 0$

**For all rewrites:** generate a rewrite-of-rewrite in the opposite direction

**ATT:**

$$\widehat{\mathrm{ATT}}_{\mathrm{RATE}} = \frac{1}{n_1} \sum_{i:\ w^i=1} \Big[ R\big(x^i, \mathrm{Re}(x^i, \mathrm{Re}(x^i, y^i, 0), 1)\big) - R\big(x^i, \mathrm{Re}(x^i, y^i, 0)\big) \Big].$$

**ATU:**

$$\widehat{\mathrm{ATU}}_{\mathrm{RATE}} = \frac{1}{n_0} \sum_{i:\ w^i=0} \Big[ R\big(x^i, \mathrm{Re}(x^i, y^i, 1)\big) - R\big(x^i, \mathrm{Re}(x^i, \mathrm{Re}(x^i, y^i, 1), 0)\big) \Big].$$

**ATE:**

$$\widehat{\mathrm{ATE}}_{\mathrm{RATE}} = \frac{n_1}{n_1 + n_0} \widehat{\mathrm{ATT}}_{\mathrm{RATE}} \ + \ \frac{n_0}{n_1 + n_0} \widehat{\mathrm{ATU}}_{\mathrm{RATE}}.$$

**return** $\widehat{\mathrm{ATT}}_{\mathrm{RATE}}$, $\widehat{\mathrm{ATU}}_{\mathrm{RATE}}$, $\widehat{\mathrm{ATE}}_{\mathrm{RATE}}$

# 5 Theoretical Analysis

We show that RATE is both unbiased and $\sqrt{n}$-consistent under mild assumptions.

## 5.1 Latent Variable Model and Assumptions

Consider representing each response $Y$ as $Y(W, Z, \xi)$, where:

- $W$ is the target binary attribute
- $Z$ are off-target attributes that remain *invariant* under rewriting
- $\xi$ are off-target attributes that *may* change inadvertently under rewriting

We then assume:

**Assumption 5.1** (Additive Reward Decomposition)**.**

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

**Assumption 5.2** (Rewrite Distribution)**.** Rewriting $Y(W, Z, \xi)$ to enforce $W = w_{\mathrm{new}}$ yields a distribution over $\xi$ values independent of $(W, Z)$. Formally,

$$\mathrm{Re}\big(X, Y(W, Z, \xi), w_{\mathrm{new}}\big) \ \stackrel{d}{=} \ Y\big(w_{\mathrm{new}}, Z, \widetilde{\xi}\big), \quad \text{where } \widetilde{\xi} \sim \mathbb{P}_{\mathrm{Re}}(\widetilde{\xi}).$$

Assumption 5.1 states that $\xi$ (mutable attributes such as stylistic artifacts) *add* to the reward in a way that does not interact with $W$ or $Z$. Assumption 5.2 posits that rewriting is essentially sampling a new $\xi$ from some distribution $\mathbb{P}_{\text{Re}}$, representing the random off-target changes that an LLM might introduce.

Under these assumptions, we have:

**Theorem 5.3** (Unbiasedness and Consistency)**.** *Let* $R(X, Y(W, Z, \xi)) = R_{W,Z}(X, W, Z) + R_\xi(X, \xi)$ *and* $Re(Y(W, Z, \xi), 1 - W) \stackrel{d}{=} Y(1 - W, Z, \tilde{\xi})$ *where* $\tilde{\xi} \stackrel{d}{\sim} P_{Re}(\tilde{\xi})$*. Also assume* $R(\cdot, \cdot)$ *is bounded. Suppose we have a set of prompt-completion pairs* $\{x^i, y^i\}$ *sampled i.i.d., with* $P(W = 1) \in (0, 1)$*. Then the RATE estimators, defined as:*

$$\widehat{ATT}_{RATE} = \frac{1}{n_1} \sum_{i:w^i=1} [R(x^i, Re(Re(y^i, 0), 1)) - R(x^i, Re(y^i, 0))]$$
$$\widehat{ATU}_{RATE} = \frac{1}{n_0} \sum_{i:w^i=0} [R(x^i, Re(y^i, 1)) - R(x^i, Re(Re(y^i, 1), 0))]$$
$$\widehat{ATE}_{RATE} = \frac{n_1}{n_0 + n_1} \widehat{ATT}_{RATE} + \frac{n_0}{n_0 + n_1} \widehat{ATU}_{RATE}$$

*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 a complete proof.

## 5.2 Intuitive Explanation of the Proof

The double-rewrite operation effectively ensures that the random off-target effects introduced in rewriting cancel out in expectation. By rewriting $W = 1$ text to $W = 0$ and then rewriting back to $W = 1$, we stay in the same "rewrite distribution space" and thus keep the off-target $v$ artifacts aligned in expectation. Consequently, the difference in rewards isolates the effect of going from $W = 0$ to $W = 1$. A similar argument applies for going from $W = 1$ to $W = 0$.

**Justifying Assumption 5.1:** 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 ($\xi$). 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 $\xi$, 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.

**Justifying Assumption 5.2:** LLMs have a characteristic way of writing that appears to consistent across contexts. For instance, LLMs will avoid using profanity or grammatical errors in most contexts. Thus, it is

plausible that the distribution of $\xi$ values introduced by rewriting is independent of the original $W, Z$ value (though this is ultimately an empirical question likely to vary across LLMs).

# 6 Empirical Evaluation

In this section, we outline how RATE is applied in practice and illustrate its efficacy on both real-world and synthetic data. We focus on a brief summary here; for more extensive results, see the original paper on which this thesis is based.

## 6.1 Implementation Details

**LLM-Based Rewriting.** We use an LLM (e.g., GPT-4) to generate rewrites in a batch manner. For each sample $(x^i, y^i, w^i)$, we:

1. For the examples with $w^i = 0$, prompt the LLM: *"Rewrite this response so that [attribute] = 1 while changing nothing else."*
2. Similarly, for the examples with $w^i = 1$, prompt: *"Rewrite this response so that [attribute] = 0 while changing nothing else."*
3. For each of these two rewrites, do a second rewrite to flip the attribute back.

**Example Prompt.** Suppose $W$ is *sentiment*, and the original text is negative. We instruct:

> *"Rewrite the following text to make it **positive in sentiment**, but do not change anything else like factual content or style."*

**Handling Edge Cases.** In some datasets, an example might already exhibit a contradictory combination of attributes (e.g., negative text that also has local positivity). We rely on the LLM's general handling, but do manually inspect a small subset of rewrites to ensure they are not totally off.

**Computational Cost.** Each rewrite is a forward pass through an LLM. For 25K examples, generating rewrites and rewrites-of-rewrites can cost tens of dollars with current APIs, which is acceptable for research-scale experiments but may be expensive at higher volume. At any rate, the cost is much lower than the cost of collecting human-generated counterfactuals.

Once the rewrites are generated, we pass them through a reward model, yielding the estimates $\widehat{\text{ATT}}_{\text{RATE}}$, $\widehat{\text{ATU}}_{\text{RATE}}$, and $\widehat{\text{ATE}}_{\text{RATE}}$.

## 6.2 Real-World Data and Observations

We apply RATE to real-world reward models and real-world data. We observe several key phenomena:

**Comparison to Naive Estimator.** We see large discrepancies between the naive estimates of the effect of $W$ vs. the RATE-based estimates. In many settings, naive methods suggest the reward model strongly

"favors" length, but RATE reveals a substantially smaller effect, indicating that length might have been correlated with truly helpful content.

**Rewrite-of-Rewrite Necessity.** We also compare a *single-rewrite* approach (comparing $y$ to $\mathrm{Re}(x, y, \bar{w})$) to the double rewrite. The single-rewrite approach consistently yields biased estimates that overstate the effect of $W$. The difference can be substantial—on some tasks, the sign of the effect even flips.

### 6.3 Synthetic Experiments

In a semi-synthetic setup, we synthesize text with known $W$ attributes (positive vs. negative, short vs. long) and artificially correlate $W$ with another attribute. We then use a black-box reward model (e.g., a sentiment classifier). Since we know the ground truth correlation structure, we can test how each estimator behaves under distributional shifts:

- **Naive Estimator** exhibits large changes in reported effect whenever the correlation is increased or decreased.
- **RATE** remains stable and aligned with the (assumed) ATE across different correlation levels, validating the causal interpretation.

Overall, these empirical studies show that RATE is both practical (easy to implement with current LLM APIs) and significantly more robust than naive alternatives.

## 7 Discussion and Limitations

**Generalization to Contrastive Rewards.** As hinted in Section 3, RATE naturally extends to $R(x, y_1, y_0)$ by looking at pairs of rewrites in the contrastive setting. The result is an estimate of

$$\mathbb{E}\big[R(X,\, Y(1),\, Y(0))\big],$$

i.e., how changing one attribute in isolation of everything else affects the model's *relative* preference.

**Model Edits and Steering.** An interesting extension is to compare two different models, $\pi$ vs. $\pi_0$, using:

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

This can reveal whether a fine-tuned or otherwise modified model truly changed its behavior on attribute $W$, again controlling for off-target shifts. This is particularly useful for evaluating "steering vectors" $\lambda$ where $\lambda$ is added to the residual stream of a transformer-based LLM, inducing the distribution $\pi_\lambda$. RATE can help determine whether $\lambda$ truly steers the model in the intended direction.

**Limitations.** Despite its benefits, RATE relies on:

- **Quality of LLM Rewrites.** If rewriting instructions are misunderstood or if the LLM is unwilling to produce certain rewrites (e.g., making text intentionally unhelpful), it can be challenging to gather valid pairs.
- **No Guarantee of Perfect Attribute Control.** We assume rewrites yield texts that differ *only* in $W$, but LLMs may drift in subtle ways.
- **Additive Decomposability.** While plausible for many aspects of human preferences, the additive form of $R$ (Assumption 5.1) is not guaranteed in every scenario.

# 8 Conclusion and Future Work

This thesis introduced **RATE (Rewrite-based Attribute Treatment Estimators)**, a method for estimating the causal effect of a particular textual attribute on a reward model's outputs. By leveraging LLMs to generate approximate counterfactuals and offsetting their imperfections via double rewriting, RATE delivers a principled and practical way to evaluate whether a reward model *truly* responds to an attribute or merely exploits spurious correlations.

**Directions for Future Research.**

- *Enhancing Rewrite Quality.* Improvements in prompt design, or specialized rewrite models, may yield even better (i.e., more truly "counterfactual") text pairs.
- *Prompt-Rewriting vs. Response-Rewriting.* Instead of rewriting the responses, one could rewrite the *prompts*, then generate responses with or without attribute $W$. This approach might remove some complexities but introduces others.
- *Beyond Binary Attributes.* Many attributes of interest are continuous or multi-class (e.g., a politeness scale). Extending RATE to these cases is an intriguing generalization.

Overall, RATE paves the way for more causally rigorous auditing of reward models, ensuring that alignment and preference models do not drift toward unintended or spurious attributes.

# A Proof of Main Theorem

**Theorem A.1** (Unbiasedness and $\sqrt{n}$-Consistency of RATE—Consistent Notation)**.** *Let $R(X, Y(W, Z, \xi)) = R_{W,Z}(X, W, Z) + R_\xi(X, \xi)$ and $Re(Y(W, Z, \xi), 1-W) \stackrel{d}{=} Y(1-W, Z, \tilde{\xi})$ where $\tilde{\xi} \stackrel{d}{\sim} P_{Re}(\tilde{\xi})$. Also assume $R(\cdot, \cdot)$ is bounded. Suppose we have a set of prompt-completion pairs $\{x^i, y^i\}$ sampled i.i.d., with $P(W = 1) \in (0, 1)$. Then the RATE estimators, defined as:*

$$\widehat{ATT}_{RATE} = \frac{1}{n_1} \sum_{i:w^i=1} [R(x^i, Re(Re(y^i, 0), 1)) - R(x^i, Re(y^i, 0))]$$
$$\widehat{ATU}_{RATE} = \frac{1}{n_0} \sum_{i:w^i=0} [R(x^i, Re(y^i, 1)) - R(x^i, Re(Re(y^i, 1), 0))]$$
$$\widehat{ATE}_{RATE} = \frac{n_1}{n_0 + n_1} \widehat{ATT}_{RATE} + \frac{n_0}{n_0 + n_1} \widehat{ATU}_{RATE}$$

*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 of Theorem A.1.* First, we'll prove the unbiasedness and $\sqrt{n}$-consistency of $\widehat{\text{ATT}}_{\text{RATE}}$. The argument for $\widehat{\text{ATU}}_{\text{RATE}}$ follows by symmetry. Then, we can use these results to prove the same for $\widehat{\text{ATE}}_{\text{RATE}}$. Throughout, we use $\tilde{\xi}$ and $\tilde{\tilde{\xi}}$ to denote i.i.d. samples from the distribution $P_{\text{Re}}$, where the former comes from the first rewrite and the latter from the rewrite of the rewrite.

**1. Unbiasedness and $\sqrt{n}$-Consistency of $\widehat{\text{ATT}}_{\text{RATE}}$** Fix a prompt $x$ and response $y$ with $w = 1$, omitting superscripts for convenience. Then by our latent variable model, $y = Y(1, z, v)$ for some realizations $z$ and $v$ of $Z$ and $\xi$. We calculate:

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

which has expected value:

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

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

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

For $\sqrt{n}$-consistency, note that $R(\cdot,\cdot)$ is bounded, so its variance is bounded. As the $x^i, y^i$ are i.i.d., so are the $R(x^i, y^i)$. Thus, $\widehat{\text{ATT}}_{\text{RATE}}$ is an average over $n_1$ i.i.d. random variables with finite variance, implying:

$$\sqrt{n_1}(\widehat{\text{ATT}}_{\text{RATE}} - \text{ATT}) = O_p(1)$$

Since $\frac{n_1}{n} \xrightarrow{p} P(W=1)$ and $P(W=1) \in (0,1)$, we have $\sqrt{\frac{n}{n_1}} = O_p(1)$, which implies:

$$\sqrt{n}(\widehat{\text{ATT}}_{\text{RATE}} - \text{ATT}) = O_p(1)$$

**2. Unbiasedness and $\sqrt{n}$-Consistency of $\widehat{\text{ATU}}_{\text{RATE}}$** By the same argument as for ATT and since $P(W=0) \in (0,1)$:

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

and

$$\sqrt{n}(\widehat{\text{ATU}}_{\text{RATE}} - \text{ATU}) = O_p(1)$$

**3. Unbiasedness and $\sqrt{n}$-Consistency of $\widehat{\text{ATE}}_{\text{RATE}}$** The ATE estimator is a weighted average of the ATT and ATU estimators. By the law of total expectation:

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

For $\sqrt{n}$-consistency, we can write:

$$\sqrt{n}(\widehat{\text{ATE}}_{\text{RATE}} - \text{ATE}) = \frac{n_1}{n}\sqrt{n}(\widehat{\text{ATT}}_{\text{RATE}} - \text{ATT}) + \frac{n_0}{n}\sqrt{n}(\widehat{\text{ATU}}_{\text{RATE}} - \text{ATU})$$

Since:

$$\frac{n_1}{n} \xrightarrow{p} P(W=1), \frac{n_0}{n} \xrightarrow{p} P(W=0)$$
$$\sqrt{n}(\widehat{\text{ATT}}_{\text{RATE}} - \text{ATT}) = O_p(1), \sqrt{n}(\widehat{\text{ATU}}_{\text{RATE}} - \text{ATU}) = O_p(1)$$

By Slutsky's theorem:

$$\sqrt{n}(\widehat{\text{ATE}}_{\text{RATE}} - \text{ATE}) = O_p(1)$$

□