Causal Evaluations of Black-Box Rewards Models via Textual Interventions

Sean Richardson

Advisor: Victor Veitch

Approved _____

Date _____

March 15, 2025

Abstract

Reward models aim to align large language models (LLMs) by approximating human preferences, but interpreting what they actually reward is nontrivial—especially given their black-box nature. The core challenge is **confounding**: textual attributes of interest (e.g., helpfulness or sentiment) are often correlated with others, such as length or style. As a result, naive comparisons—like differences in average reward across attribute groups—fail to isolate causal effects. Standard causal methods, such as randomized interventions or adjustment for confounders, also break down: we cannot perfectly intervene on individual attributes without affecting others, and we cannot feasibly enumerate or control for all confounding features in text.

This thesis introduces **RATE (Rewrite-based Attribute Treatment Estimators)**, a causal inference framework tailored to these challenges. RATE uses large language models to generate approximate interventions and introduces a novel double-rewrite strategy to cancel out unintended, off-target changes introduced during rewriting. Under mild assumptions, we show that RATE yields unbiased, $\sqrt{n}$-consistent estimates of treatment effects (ATE, ATT, ATU), despite relying only on imperfect interventions and black-box access to the reward model. Empirically, RATE uncovers spurious correlations that naive methods overlook, offering a principled and practical approach to interpreting reward 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.

Contents

1 Introduction

1.1 Motivation

In the context of large language models (LLMs), reward models evaluate the quality 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 common approach is to train a reward model based on human pairwise preferences for which of two responses is "more helpful." This supervised learning process does not guarantee that the reward model will learn to value helpfulness per se. Instead, it may learn to rely on other attributes that happen to correlate with helpfulness (e.g., length, style, politeness), so long as this correlation allows for good performance on the training data.

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. If we were to use this reward model to fine-tune a language model, we might inadvertently encourage the model to generate longer responses without actually increasing helpfulness. To address this challenge, we need explainability methods to understand what the reward model is actually rewarding.

A Naive Approach. One might try to isolate the impact of an attribute W by simply partitioning the data into $W = 1$ and $W = 0$ examples and comparing mean reward. However, if W is correlated with other features of the text, that naive approach will conflate the effect of W with these other features. For example, if helpful responses tend to be longer, the naive approach would suggest that helpfulness drives reward, even supposing that the reward model is actually indifferent to helpfulness but merely prefers longer responses.

1.2 Our Contribution

Conceptually, what we care about is the effect of changing one specific text attribute while keeping other text aspects fixed. If the reward model is actually indifferent to the attribute, then this effect should be zero. If the reward model is actually sensitive to the attribute, then this effect will be non-zero. This is a causal inference problem, where the treatment is the attribute W and the outcome is the reward.

To properly estimate how the reward would change if we were to *intervene* on one specific text attribute while keeping other text aspects fixed, we adopt a causal inference perspective. We formalize the problem as a treatment effect estimation problem, where the treatment is the attribute W and the outcome is the reward.

If we were able to perfectly intervene on one specific text attribute while keeping other text aspects fixed, causal estimation would be straightforward. However, scalability requires the use of automated methods, and LLM-based rewrites are typically *imperfect*. LLMs may inadvertently alter additional attributes, for instance by changing the formatting or style of the text. Our key technical insight is that by rewriting *twice* and comparing the resulting reward differences between the two rewrites, we can correct for these off-target changes. We prove that 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).

1.3 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. 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 [4]. *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 [9] create diverse counterfactuals, but often do not explicitly correct for rewriting artifacts. CEBaB [1] 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.** [6] propose large-scale benchmarks for reward models without taking a causal perspective. [7] 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.

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 analysis can be extended to the contrastive case, but we focus on the pointwise version for simplicity. A reward model may be prespecified, in which case we do not need to rely on explainability methods. However, in many cases, the reward model is a complex, opaque function (e.g., a neural network) that we cannot easily interpret. In such cases, we need explainability methods in order to understand what the model is actually rewarding.

Our key motivating application is reinforcement learning from human preferences [2], where the reward model is trained to predict human preferences from pairwise comparisons. In this framework, 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 goal is to infer some general structure that explains the preferences. We posit a statistical model according to which these preferences are driven by some oracle reward function mapping prompt-response pairs to scalar values, $R^*(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. That is, we model $\Pr(y_1 > y_0) = \phi(R^*(x, y_1), R^*(x, y_0))$ where ϕ is some function mapping pointwise rewards to preference probabilities.

The Bradley-Terry model is a common choice for $\Pr(y_1 > y_0)$, using $\phi(R^(x, y_1), R^(x, y_0)) = \sigma(R^(x, y_1) - R^(x, y_0))$ where σ is the logistic function. A parametric reward model $R_\theta(x, y)$ is then trained to predict $\Pr(y_1 > y_0)$ by maximum likelihood estimation:

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

3.2 Notation and Causal Framework

Let $(x^i, y^i)_{i=1}^n$ be a dataset of n prompt-response pairs, where x^i represents a prompt and y^i is the corresponding response. For each pair, we observe a reward $R(x^i, y^i) \in \mathbb{R}$ assigned by the reward model.

We focus on a binary attribute $W(x^i, y^i) \in \{0, 1\}$ that potentially affects the reward. Examples include:

- Whether a response is "helpful" versus "unhelpful"
- Whether a response has "positive" versus "negative" sentiment
- Whether a response is "long" versus "short"

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:

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

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

Unfortunately, $Y(1)$ 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 [5].

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 [3]. That is, we identify some sufficient adjustment set Z and estimate the ATE as:

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

However, in high-dimensional text settings, this approach faces challenges:

- **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, 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.

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:

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

where $W(x^i, \tilde{y}^i) = w_{\text{new}}$. That is, $\tilde{y}^i$ is the LLM-generated rewrite of y^i forcing $W(\tilde{y}^i) = w_{\text{new}}$ but aims to otherwise leave the text unchanged. 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 Solution

Challenge: LLM rewrites can inadvertently alter off-target attributes (e.g., grammar, length, writing style) beyond the target attribute W. A naive comparison $R(x, y) - R(x, \mathrm{Re}(x, y, w_{\text{new}}))$ would capture both the effect of changing W and these unintended modifications.

Solution: The Double-Rewrite Technique

Instead of comparing an original text to a single rewrite, RATE uses a two-step rewriting process:

1. **First rewrite:** From original attribute value to the opposite

2. **Second rewrite:** From the opposite back to the original

This process creates a "rewrite-of-rewrite" that, we hope, shares the same rewriting artifacts as the first rewrite but has the original attribute value. By comparing these two texts, we isolate the effect of the attribute change.

For examples where $W = 1$ originally:

1. Start with y where $W = 1$
2. Create first rewrite: $\tilde{y} = \mathrm{Re}(x, y, 0)$ where $W = 0$
3. Create second rewrite: $\tilde{\tilde{y}} = \mathrm{Re}(x, \tilde{y}, 1)$ where $W = 1$ again
4. Compute difference: $R(x, \tilde{\tilde{y}}) - R(x, \tilde{y})$

For examples where $W = 0$ originally:

1. Start with y where $W = 0$
2. Create first rewrite: $\tilde{y} = \mathrm{Re}(x, y, 1)$ where $W = 1$
3. Create second rewrite: $\tilde{\tilde{y}} = \mathrm{Re}(x, \tilde{y}, 0)$ where $W = 0$ again
4. Compute difference: $R(x, \tilde{y}) - R(x, \tilde{\tilde{y}})$

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

5.1 Latent Variable Model and Assumptions

Earlier we motivated the double-rewrite technique by noting that it creates a "rewrite-of-rewrite" that, we hope, shares the same rewrite artifacts as the first rewrite but has the original attribute value. It turns out that causal inference only requires a weaker form of this, namely that the rewrite artifacts introduced in steps 2 and 3 will be similar in distribution, allowing them to cancel out in expectation.
To make this more precise, we represent each response Y as $Y(W, Z, \xi)$, where:

- W is the target binary attribute (e.g., helpfulness)
- Z are off-target attributes that remain *invariant* under rewriting (e.g., topic, language)
- ξ are off-target attributes that *may* change inadvertently under rewriting (e.g., style, length)

Now we can state precisely the assumptions we need to make about the rewriting process:

Assumption 5.1 (Rewrite Distribution)**.** Rewriting $Y(W, Z, \xi =)$ to enforce $W = w_{\text{new}}$ replaces ξ with $\tilde{\xi}$ drawn from a distribution independent of (W, ξ). That is,

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

Note that this assumption is not as strong as it may seem. It does not mean that the mutable off-target attributes must remain the same across the rewrite and rewrite-of-rewrite, only that they must be drawn from the same distribution. (In fact, this assumption can be weakened to the assumption that they are drawn from distributions with the same expectation, but we do not pursue this here.)

We also need to make some assumptions about how the reward model handles these latent variables. In particular, we assume that the reward model lacks any interaction effects between W and ξ or Z and ξ.

Assumption 5.2 (Additive Reward Decomposition)**.**

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

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* $Re(Y(W, Z, \xi), 1 - W) \overset{d}{=} Y(1 - W, Z, \tilde{\xi})$ *where* $\tilde{\xi} \overset{d}{\sim} P_{Re}(\tilde{\xi})$*. 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 key insight behind RATE is that we can isolate the causal effect of attribute W by accounting for both:

1. **The intended change:** The shift from $W = 0$ to $W = 1$ (or vice versa)
2. **The unintended changes:** Off-target attribute modifications introduced by the rewriting process

Under our additive decomposition assumption, the reward model's output can be separated into two components:

- $R_{W,Z}(X, W, Z)$: The component affected by our target attribute W and immutable attributes Z
- $R_\xi(X, \xi)$: The component affected by mutable artifacts ξ that may change during rewriting

When we rewrite a text, the LLM samples a new ξ from some distribution $\mathbb{P}_{\text{Re}}$. The double-rewrite technique ensures that both the first rewrite and the rewrite-of-rewrite have ξ values sampled from the same distribution. In expectation, their contributions to the reward (R_ξ) cancel out, leaving only the effect of changing W.

Mathematically, if we start with $y = Y(1, Z, \xi)$ and create $\tilde{y} = Y(0, Z, \tilde{\xi})$ and $\tilde{\tilde{y}} = Y(1, Z, \tilde{\tilde{\xi}})$, then:

$$\mathbb{E}[R(x, \tilde{\tilde{y}}) - R(x, \tilde{y})] = \mathbb{E}[R_{W,Z}(x, 1, Z) - R_{W,Z}(x, 0, Z)] = \text{ATT}$$

This captures precisely the causal effect we seek to measure.

Justifying Assumption 5.2: 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.

Justifying Assumption 5.1: 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 ξ 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 [8].

6.1 Implementation Details

LLM-Based Rewriting. We implement RATE using modern LLMs (namely, GPT-4) with carefully crafted prompts. For each sample $(x^i, y^i, W(x^i, y^i))$, we follow a structured rewriting protocol. The original paper provides more details on the rewriting process, but we summarize it here for convenience:

1. **First rewrite generation:**
   - For examples with $W(x^i, y^i) = 0$, we prompt: *"Rewrite this response so that [attribute] = 1 while preserving all other aspects."*
   - For examples with $W(x^i, y^i) = 1$, we prompt: *"Rewrite this response so that [attribute] = 0 while preserving all other aspects."*
2. **Second rewrite generation:** For each first rewrite, we generate a second rewrite that reverts the attribute to its original value.
3. **Reward calculation:** We compute the reward for all rewrites and calculate the difference according to the RATE equations.

Prompt Engineering. The quality of rewrites is crucial for RATE's effectiveness. We design our prompts to:

1. Clearly specify the target attribute change
2. Explicitly instruct preservation of other textual properties
3. Provide concrete examples when necessary

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."*

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). This is inevitable when dealing with murky concepts like "helpfulness" or "sentiment." 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 $\widehat{\mathrm{ATT}}_{\mathrm{RATE}}$, $\widehat{\mathrm{ATU}}_{\mathrm{RATE}}$, and $\widehat{\mathrm{ATE}}_{\mathrm{RATE}}$.

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 other rewarded attributes.

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 yields very different estimates, indicating the importance of the double-rewrite technique.

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.
- **Single-Rewrite Estimator** exhibits more instability, indicating the importance of the double-rewrite technique.

7 Discussion and Limitations

7.1 Key Contributions Revisited

This thesis makes three principal contributions to the evaluation of reward models:

1. **A formal causal framework** for interpreting reward models

2. **The RATE methodology** that leverages LLMs to generate approximate counterfactuals and correct for rewriting artifacts

3. **Theoretical guarantees** of unbiasedness and $\sqrt{n}$-consistency under mild assumptions

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, π vs. π_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" λ where λ is added to the residual stream of a transformer-based LLM, inducing the distribution π_λ. RATE can help determine whether λ truly steers the model in the intended direction.

Limitations and Future Work. 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), we may not be capturing the causal effect of interest.

- **Additive Decomposability.** While plausible, the additive form of the reward model (Assumption 5.2) is not guaranteed in every scenario.

Future work could address these limitations and extend RATE in several directions:

- *Enhancing Rewrite Quality.* Improvements in prompt design, or specialized rewrite models, may yield even better rewrites.

- *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 could be valuable.

8 Conclusion

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 (in expectation) their imperfections via double rewriting, RATE delivers a principled and practical way to evaluate whether a reward model *really* rewards an attribute or merely a correlated attribute.

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 $Re(Y(W, Z, \xi), 1-W) \overset{d}{=} Y(1-W, Z, \tilde{\xi})$ where $\tilde{\xi} \overset{d}{\sim} P_{Re}(\tilde{\xi})$. 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_{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 ξ. 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:

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

□

References

[1] E. D. Abraham, K. D'Oosterlinck, A. Feder, Y. Gat, A. Geiger, C. Potts, R. Reichart, and Z. Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems*, 35:17582–17596, 2022.

[2] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[3] C. Cinelli, A. Forney, and J. Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104, 2024.

[4] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, B. M. Stewart, V. Veitch, and D. Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond, 2022.

[5] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[6] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.

[7] J. Park, S. Jwa, M. Ren, D. Kim, and S. Choi. Offsetbias: Leveraging debiased data for tuning evaluators, 2024.

[8] D. Reber, S. Richardson, T. Nief, C. Garbacea, and V. Veitch. Rate: Causal explainability of reward models with imperfect counterfactuals, 2025.

[9] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models, 2021.