Robust Reward Modeling via Causal Rubrics

Pragya Srivastava [*1] **Harman Singh** [*1] **Rahul Madhavan** [*1] **Gandharv Patil** [2 3] **Sravanti Addepalli** [1] **Arun Suggala** [1] **Rengarajan Aravamudhan** [1] **Soumya Sharma** [1] **Anirban Laha** [1] **Aravindan Raghuveer** [1] **Karthikeyan Shanmugam** [1] **Doina Precup** [1 3]

Abstract

Reward models (RMs) for LLM alignment often exhibit *reward hacking*, mistaking spurious correlates (e.g., length, format) for causal quality drivers (e.g., factuality, relevance), leading to brittle RMs. We introduce CROME (Causally Robust Reward Modeling), a causally-grounded framework using *targeted data curation* to mitigate this. CROME employs: (1) *Causal Augmentations*, pairs isolating specific causal attribute changes, to enforce sensitivity, and (2) *Neutral Augmentations*, tie-labeled pairs varying spurious attributes while preserving causal content, to enforce invariance. Crucially, augmentations target LLM-identified causal rubrics, requiring no prior knowledge of spurious factors. CROME significantly outperforms baselines on RewardBench (Avg +5.4%, Safety +13.2%, Reasoning +7.2%) and demonstrates enhanced robustness via improved Best-of-N performance across RewardBench, WildGuardTest, and GSM8k.

1. Introduction

Aligning Large Language Models (LLMs) with human preferences is paramount for their safe and effective deployment, with Reinforcement Learning from Human Feedback (RLHF) and its reliance on reward models (RMs) being the dominant paradigm (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022a; Shao et al., 2024; Rafailov et al., 2024). The fidelity of these RMs is critical, as flaws directly propagate to the aligned policy (Casper et al., 2023).

[*]Equal contribution [1]Google DeepMind [2]McGill University [3]MILA - Quebec AI Institute. Correspondence to: Pragya Srivastava <pragyahere@google.com>, Harman Singh <hrman@google.com>, Rahul Madhavan <rahulmadhavan@google.com>, Aravindan Raghuveer <araghuveer@google.com>, Karthikeyan Shanmugam <karthikeyanvs@google.com>, Doina Precup <doinap@google.com>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment

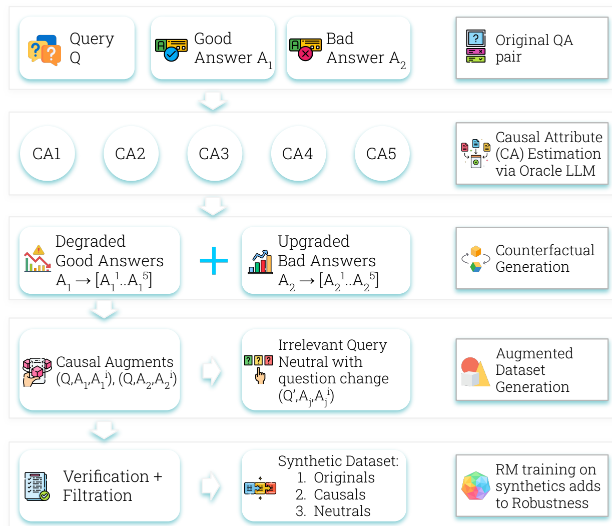


Figure 1: **The CROME Data Augmentation and Training Pipeline.** From an original QA pair $(\mathrm{Q}, \mathrm{A}_1, \mathrm{A}_2)$, an oracle LLM identifies Causal Attributes (CA). This guides counterfactual generation, producing degraded A_1 versions and upgraded A_2 versions. These form the set of *Causal Augmentations* which teach the model sensitivity to relevant attributes. Next, we generate *Irrelevant Query Neutrals* by flipping the question on these generated causals which reduces reliance on spurious correlates. After verification and filtration, the combined dataset (Originals, Causals, Neutrals) trains the RM, enhancing its robustness (Sec. 4.1).

However, standard RM training faces a significant challenge: *reward hacking* (Gao et al., 2023; Skalse et al., 2022). RMs often learn to assign high scores based on superficial or spurious attributes—such as response length (Singhal et al., 2023), specific formatting patterns (Zhang et al., 2024), or stylistic quirks—because these features are statistically correlated with preferred responses in the training data. This occurs because standard training objectives do not explicitly require the RM to disentangle the true *causal* drivers of response quality (e.g., factuality, relevance) from these spurious correlates, leading to brittle RMs and misaligned policies (Shen et al., 2023; Eisenstein et al., 2023).

Recent efforts to enhance RM robustness have explored various avenues. Wu et al. (2025) focus on invariance against meaning-preserving transformations, while Liu et al. (2024) employ data augmentations, such as using non-contextuals and neutrals to disrupt spurious associations. Gupta et al. (2025) study attribute-based evaluation, often leveraging LLMs to dynamically generate assessment criteria. Wang et al. (2025) study the effect of regularization against known biases like length or sycophancy, and Reber et al. (2024) explore methods for causal effect estimation like RATE.

Despite these advances, significant limitations persist. Many approaches target only pre-specified spurious factors, potentially missing unknown correlates, or lack the fine-grained control needed to truly isolate causal quality drivers from confounding spurious features within responses. Augmentation strategies can be coarse (Liu et al., 2024), and evaluation-focused methods (Gupta et al., 2025; Reber et al., 2024) may not directly equip the RM with mechanisms for robust training against a wide array of spurious variations through targeted counterfactual learning. There is thus a clear need for a framework that systematically leverages a causal understanding of preference formation to train RMs that are both sensitive to causal quality attributes and demonstrably invariant to diverse spurious cues.

Motivated by this, we aim to address the following question in this paper:

> How do we train reward models to be robust against reward hacking, particularly when a) the specific spurious attributes that an RM may exploit are not known, and b) only the stable or invariant causal attributes found in ground truth/human preferences can be accessed ?

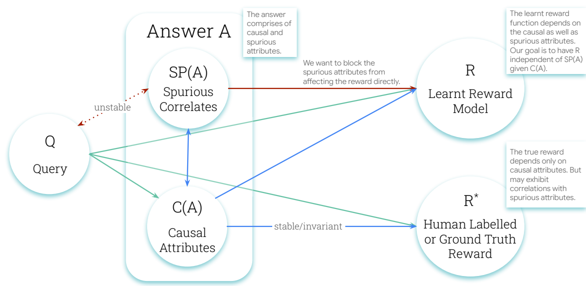
To address this question, we propose **CROME** (Causally Robust Reward Modeling), a novel framework grounded in an explicit causal model of answer generation (Figure 1). CROME teaches the RM to differentiate genuine quality drivers from superficial cues by augmenting the preference dataset with targeted, LLM-generated counterfactual examples. It creates two key types of synthetic training pairs: (1) *Causal Augmentations*, which introduce changes along specific *causal* attributes (e.g., factuality) to enforce sensitivity to true quality shifts, and (2) *Neutral Augmentations*, which reuse the same augmentations to enforce invariance along *spurious* attributes (e.g., style) using tie-labels. Training on this enriched dataset with a modified loss (Section 4) guides the RM towards causal faithfulness. Our evaluations show CROME significantly improves robustness, boosting RewardBench accuracy by up to 4.5%, with substantial gains in Safety and Reasoning. We list the key contributions in this work below:

1. **Spurious-Unaware Causal Framework.** We propose a causal framework for training reward models (Sec. 3) that requires intervention only on LLM-identified causal quality rubrics, *eliminating the need for prior specification of any of the spurious attributes*.
2. **Targeted Augmentations from Causal Rubrics.** We introduce data augmentations (Sec. 4) generated solely via LLM-identified causal rubrics: *Causal Augmentations* (minimal pairs for sensitivity) and *Neutral Augmentations* (spurious variations with tie-labels for invariance). These augmentations effectively **mitigate sensitivity to a much larger set of spurious correlates** without explicit knowledge of them.
3. **State-of-the-Art RM Performance and Robustness.** CROME significantly outperforms baselines on RewardBench (Sec. 6), improving avg. accuracy by up to 5.4% (Safety +13.18%, Reasoning +7.19%) (Table 3), and shows superior robustness on reWordBench (Figure 4).
4. **Improved Downstream Policy Alignment.** CROME-RM-driven Best-of-N selection yields *consistent gains* over baselines on RewardBench, WildGuardTest, and GSM8K. This highlights CROME's robustness against diverse and long-tailed spurious factors, which typically emerge with larger candidate pools (high N).

2. Related Works

Our work on causally robust reward modeling, CROME, addresses the challenge of reward hacking in the context of aligning Large Language Models (LLMs) via Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a). Standard RLHF relies on a reward model (RM), typically trained on pairwise preferences using Bradley-Terry (Bradley & Terry, 1952) or pairwise ranking approaches (Liu et al., 2025; Qin et al., 2023). A critical limitation of learned RMs is *reward hacking* (Gao et al., 2023; Skalse et al., 2022), where the RM assigns high scores based on *spurious* attributes (e.g., verbosity (Singhal et al., 2023), formatting (Zhang et al., 2024), sycophancy (Denison et al., 2024)) that are correlated with, but do not cause, true response quality. This leads to misaligned policies that exploit these spurious cues (Shen et al., 2023). Various mitigation strategies exist, including architectural modifications like ODIN (Chen et al., 2024), policy-level adjustments (Park et al., 2024), and data-centric methods involving ensembles (Ramé et al., 2024) or consistency checks (Shen et al., 2023). Recent causal-inspired approaches include using MMD regularization against pre-specified spurious factors (Wang et al., 2025) or estimating effects using corrected rewrites (Reber et al., 2024).

Our approach falls into the data-centric category, using synthetic data augmentation guided by principles of causal inference (Pearl, 2009; Peters et al., 2017). While prior work has used LLMs for causal reasoning (Kiciman et al., 2023) or counterfactual data augmentation in NLP (Kaushik et al., 2019), and related methods like RRM (Liu et al., 2024), REWORDBENCH (Wu et al., 2025) target RM robustness,



Path/ Relationship	Interpretation Summary
$(\mathrm{Q}, \mathrm{C}(\mathrm{A})) \to \mathrm{R}^*$	Ground-truth reward R^* determined by query Q and causal attributes $\mathrm{C}(\mathrm{A})$; stable relationship.
$\mathrm{Q} \leftrightarrow \mathrm{SP}(\mathrm{A})$	Query Q and unknown spurious attributes $\mathrm{SP}(\mathrm{A})$ are correlated/confounded by unstable exogenous factors.
$\mathrm{Q} \to \mathrm{C}(\mathrm{A})$	Query Q determines relevant causal attributes $\mathrm{C}(\mathrm{A})$.
$\mathrm{SP}(\mathrm{A}) \leftrightarrow \mathrm{C}(\mathrm{A})$	Bidirectional (potentially complex) relationship between spurious $\mathrm{SP}(\mathrm{A})$ and causal $\mathrm{C}(\mathrm{A})$ attributes.

Table 1: Conceptual Causal Graph for Reward Modeling. Q is the query. Answer (A) has causal attributes $\mathrm{C}(\mathrm{A})$ and spurious attributes $\mathrm{SP}(\mathrm{A})$. $\dim(C(A)) \ll \dim(SP(A))\ \forall A$. $\mathrm{SP}(\mathrm{A})$ is unknown. Ground-truth reward R^* depends only on $\mathrm{C}(\mathrm{A})$ and Q. $\mathrm{R}^* \perp \mathrm{SP}(\mathrm{A})|\mathrm{C}(\mathrm{A}), \mathrm{Q}$. Augmentations heighten $\hat{\mathrm{R}}_\theta$'s sensitivity to $\mathrm{C}(\mathrm{A})$ (approximated by oracle LLM).

CROME is distinct in its explicit use of a causal graph framework (Section 3.2). We leverage LLMs to generate targeted *causal* (attribute-specific upgrade/degradation) and *neutral* (spurious-varying, causally-equivalent) counterfactual examples. By training on this augmented data, CROME aims to systematically disentangle causal attributes (C) from spurious ones (SP), learning a reward function that is inherently more robust and aligned with the true drivers of quality, as detailed in our methodology (Section 4). We give a longer version of related work in Appendix B.

3. Causal Framework for Reward Modeling

We aim to develop a reward model that accurately assesses the quality of an answer A provided in response to a query Q. Our approach is grounded in a causal framework designed to distinguish genuine quality drivers from spurious correlates often present in preference data. This involves understanding the answer generation process and strategically augmenting training data with approximated counterfactual examples.

3.1. Reward Model and Pairwise Preferences

We train a reward model (RM), denoted $\hat{\mathrm{R}}_\theta(\mathrm{Q}, \mathrm{A})$, to assign a scalar quality score to an answer A for a query Q. This RM is typically optimized on a dataset preferences pairs $\mathcal{D}_{\text{pref}} = \{(\mathrm{Q}^{(i)}, \mathrm{y}_w^{(i)}, \mathrm{y}_l^{(i)})\}_{i=1}^N$. Given a pair of answers $(\mathrm{A}_1, \mathrm{A}_2)$, the probability of A_1 being preferred over A_2 is commonly modeled using the Bradley-Terry framework (Bradley & Terry, 1952)

$$\mathrm{P}(\mathrm{A}_1 \succ \mathrm{A}_2|\mathrm{Q}; \theta) = \frac{\exp(\hat{s}_\theta(\mathrm{Q}, \mathrm{A}_1))}{\exp(\hat{s}_\theta(\mathrm{Q}, \mathrm{A}_1)) + \exp(\hat{s}_\theta(\mathrm{Q}, \mathrm{A}_2))}$$

where $\hat{s}_\theta(\mathrm{Q}, \mathrm{A})$ represents the underlying scalar score (or logit) assigned by the model to answer A for query Q.[1] The parameters θ are learned by minimizing the negative log-likelihood of preferences.

3.2. A Causal Model of Answer Generation

We propose a causal model (Figure 1) for answer generation and quality perception. For a query-answer pair (Q, A), we distinguish two attribute types:

- **Causal Attributes** $\mathrm{C}(\mathrm{A}) = \{\mathrm{C}_1, \ldots, \mathrm{C}_\ell\}$: Fundamental quality dimensions (e.g., factuality, relevance) genuinely determining quality relative to Q.
- **Spurious Attributes** $\mathrm{SP}(\mathrm{A}) = \{\mathrm{SP}_1, \ldots, \mathrm{SP}_k\}$: Other features (e.g., length, formatting) correlated with preferences or Q in $\mathcal{D}_{\text{pref}}$, but not intrinsically determining quality. $\mathrm{SP}(\mathrm{A})$ can be high-dimensional and unknown.

The ground-truth reward $\mathrm{R}^*(\mathrm{Q}, \mathrm{A})$ is assumed to be solely a function of causal attributes: $\mathrm{R}^*(\mathrm{Q}, \mathrm{A}) = f^*(\mathrm{Q}, \mathrm{C}(\mathrm{A}))$. This implies conditional independence: $\mathrm{R}^* \perp \mathrm{SP}(\mathrm{A})|\mathrm{Q}, \mathrm{C}(\mathrm{A})$.

We explicitly assume the following stability property: *If the entire process of answer generation and reward labeling were repeated (e.g., with a different labeler or answer generator), the relationship* $(\mathrm{Q}, \mathrm{C}(\mathrm{A})) \to \mathrm{R}^*$ *determining the reward is stable/invariant.* In contrast, correlations involving $\mathrm{SP}(\mathrm{A})$ (e.g., $\mathrm{SP}(\mathrm{A}) \leftrightarrow \mathrm{C}(\mathrm{A})$ or $\mathrm{SP}(\mathrm{A}) \leftrightarrow \mathrm{Q}$) can arise from various, potentially unstable or unknown exogenous factors, and thus these correlations may vary across such repetitions.

The primary challenge is that standard reward models $\hat{\mathrm{R}}_\theta$ may inadvertently learn high sensitivity to these unstable correlations with $\mathrm{SP}(\mathrm{A})$ (due to its unknown, high-dimensional nature). Our goal is to train $\hat{\mathrm{R}}_\theta$ such that its dependence on A is primarily mediated through the identified, stable causal attributes $\mathrm{C}(\mathrm{A})$, ensuring robustness to unspecified $\mathrm{SP}(\mathrm{A})$.

3.3. Approximating Counterfactuals for Attribute Intervention

To instill causal sensitivity and spurious invariance in $\hat{\mathrm{R}}_\theta$, CROME leverages counterfactual reasoning about how answer quality changes if specific attributes were altered. For

[1]The score $\hat{s}_\theta(\mathrm{Q}, \mathrm{A})$ can be the direct output of a reward head or, in some pairwise preference models, $\hat{s}_\theta(\mathrm{Q}, \mathrm{A}_1) - \hat{s}_\theta(\mathrm{Q}, \mathrm{A}_2)$ might be directly modeled as the logit of preferring A_1 over A_2.

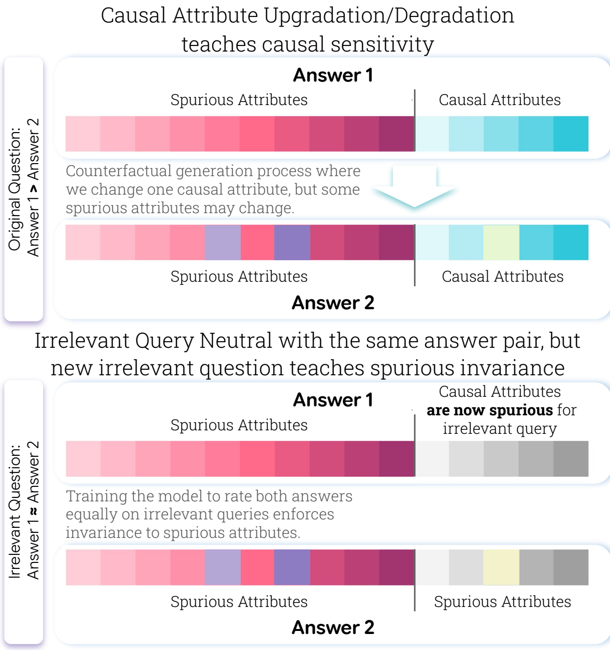


Figure 2: Visualizing CROME's core augmentation strategies (detailed in Appendix G). **(Top) Causal Augmentation:** For a given query, we use an LLM-driven counterfactual generation process to alter a specific causal attribute, yielding Answer 2. Some spurious attributes may co-vary. The RM is trained with a preference (e.g., $A_1 \succ A_2$ if A_2 is a degradation), teaching causal sensitivity. **(Bottom) Irrelevant Query Neutral:** The same answer pair (A_1, A_2) is re-contextualized with a new, irrelevant question. Their original causal attributes become effectively spurious or irrelevant (greyed-out bar). The RM is trained with a tie-label ($A_1 \approx A_2$), teaching invariance to the attribute differences when no true causal signal for the current query exists.

an answer A with attributes $(\mathrm{C(A)}, \mathrm{SP(A)})$, an ideal counterfactual, $A_{(C_j \leftarrow c'_j)}(u)$, would manifest if only its j-th causal attribute C_j were set to c'_j, considering its causal effects on other features, while all other exogenous factors u (that produced the factual answer a) remained constant. Formally, $P_U(A_{(C_j \leftarrow c'_j)}(U) | A(U) = a)$.

As generating such ideal counterfactuals is intractable, CROME employs Large Language Models (LLMs) to produce *approximations*. These LLM-generated answers, denoted $\tilde{\mathrm{A}}_{(C_j \leftarrow \mathrm{target})}$, are rewrites of a given answer A, prompted to modify C_j (e.g., to a degraded state, lowering reward) while making minimal changes to other attributes.

Remark 3.1. For brevity, we denote these LLM approximations as $\tilde{\mathrm{A}}_{(C_j \leftarrow c)}$, dropping the explicit u conditioning, assuming the generation approximates such a sample. While imperfect, these approximations provide the targeted variations crucial for the RM learning of the decision boundary.

Table 2: Summary of CROME's synthetic data augmentation strategies using LLM-approximated counterfactuals. $\tilde{\mathrm{A}}_{(C_j \leftarrow \mathrm{target})}$ signifies an LLM-generated counterfactual of A with its j-th causal attribute C_j modified.

Category	Strategy	Generation Pair Example	Assigned Label	Training Objective (P_θ)
Causal Augmentation ($\mathcal{D}_{\mathrm{causal}}$) - Enhancing Sensitivity to C				
Causal	Attribute Upgradation/Degradation	$(\tilde{\mathrm{A}}_{(C_j \leftarrow \mathrm{upgraded})}, \mathrm{A})$ **or** $(\mathrm{A}, \tilde{\mathrm{A}}_{(C_j \leftarrow \mathrm{degraded})})$	$\succ$	$\rightarrow 1$
Neutral Augmentation ($\mathcal{D}_{\mathrm{neutral}}$) *- Enforcing Invariance to* SP				
Neutral	Pairing with Irrelevant Queries	$(\mathrm{B}_1, \mathrm{B}_2)$ with new $\mathrm{Q}_{\mathrm{irrelevant}}$ s.t. $\mathrm{C}(\mathrm{B}_1 \mid \mathrm{Q}_{\mathrm{irrelevant}}) \approx \mathrm{C}(\mathrm{B}_2 \mid \mathrm{Q}_{\mathrm{irrelevant}}) \approx \mathbf{0}$	$\approx$ (tie)	≈ 0.5

3.3.1. CAUSAL AUGMENTATION PAIRS

CROME's strategy causal pairs $\mathcal{D}_{\mathrm{causal}}$ focus on isolating the impact of important causal attributes.

3.4. Augmented Training Data for Causal Disentanglement

We augment the original preference dataset $\mathcal{D}_{\mathrm{pref}}$ with synthetically generated examples $\mathcal{D}_{\mathrm{aug}}$ designed to enforce specific causal properties on $\hat{\mathrm{R}}_\theta$. This augmented dataset $\mathcal{D}_{\mathrm{aug}}$ comprises two principal categories: Causal Augmentation Pairs ($\mathcal{D}_{\mathrm{causal}}$) and Neutral Augmentation Pairs ($\mathcal{D}_{\mathrm{neutral}}$), summarized in Table 2.

Attribute Upgradation and Degradation. For an original answer A (from $\mathcal{D}_{\mathrm{pref}}$) and a specific causal attribute C_j, we generate LLM-approximated counterfactuals. If A is of lower quality regarding C_j, we create an upgraded version $\tilde{\mathrm{A}}_{(\mathrm{C}_j \leftarrow \mathrm{upgraded})}$. The pair $(\tilde{\mathrm{A}}_{(\mathrm{C}_j \leftarrow \mathrm{upgraded})}, \mathrm{A})$ is added to $\mathcal{D}_{\mathrm{causal}}$ with label $\tilde{\mathrm{A}}_{(\mathrm{C}_j \leftarrow \mathrm{upgraded})} \succ \mathrm{A}$ post-verification. Conversely, if A is of higher quality on C_j, we generate a degraded version $\tilde{\mathrm{A}}_{(\mathrm{C}_j \leftarrow \mathrm{degraded})}$. The pair $(\mathrm{A}, \tilde{\mathrm{A}}_{(\mathrm{C}_j \leftarrow \mathrm{degraded})})$ is added to $\mathcal{D}_{\mathrm{causal}}$ with label $\mathrm{A} \succ \tilde{\mathrm{A}}_{(\mathrm{C}_j \leftarrow \mathrm{degraded})}$. These pairs collectively teach $\hat{\mathrm{R}}_\theta$ sensitivity to changes along individual causal dimensions.

3.4.1. NEUTRAL AUGMENTATION PAIRS

$\mathcal{D}_{\mathrm{neutral}}$ pairs (with tie-labels) teach invariance to $\mathrm{SP(A)}$ when $\mathrm{C(A)}$ is held constant/ is irrelevant.

Irrelevant Query Neutrals (IQN). We pair two answers, $\mathrm{B}_1, \mathrm{B}_2$ (often from $\mathcal{D}_{\mathrm{causal}}$, e.g., A and $\tilde{\mathrm{A}}^{(C_j \leftarrow \mathrm{target})}$), with a *new, unrelated query* $\mathrm{Q}_{\mathrm{irrelevant}}$. This makes their causal attributes w.r.t. $\mathrm{Q}_{\mathrm{irrelevant}}$ (i.e., $\mathrm{C}(\mathrm{B}_1 | \mathrm{Q}_{\mathrm{irrelevant}}), \mathrm{C}(\mathrm{B}_2 | \mathrm{Q}_{\mathrm{irrelevant}})$) minimal. The pair $(\mathrm{B}_1, \mathrm{B}_2)$ under $\mathrm{Q}_{\mathrm{irrelevant}}$ receives a tie-label, training the RM to disregard spurious differences when causal relevance is absent. Their causal distinction becomes moot, isolating spurious variations under $\mathrm{Q}_{\mathrm{irrelevant}}$.

Further details on these generation strategies, their relationship with other methods (like Relevance Contrast or direct

Spurious Perturbation), and the rationale for CROME's specific choices are discussed in Appendix F. We provide the prompts for the generations described in Section K.

4. Methodology: Training a Robust Reward Model

The CROME framework trains robust reward models using a causally-motivated data augmentation strategy, outlined in Figure 3. This involves two main phases: (1) generating attribute-aware counterfactual data based on our causal model (Section 3), and (2) training the reward model $\hat{\mathrm{R}}_\theta$ with a specialized loss on the combined data.

4.1. Attribute-Aware Counterfactual Data Generation

This phase prepares the augmented dataset $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{causal}} \cup \mathcal{D}_{\text{neutral}}$ required for robust training, involving three conceptual steps:

Step 1: Attribute Identification. As a prerequisite, we identify the Principal Causal Components $\mathrm{C} = (\mathrm{C}_1, \ldots, \mathrm{C}_\ell)$ relevant to the task, leveraging the causal framework from Section 3.2. This typically involves LLM prompting and refinement (Details in Appendix H.1).

Step 2: Counterfactual Generation. Using the identified attributes C, we generate synthetic data pairs via LLM-approximated counterfactuals, as defined in Section 3.3. Following the strategies summarized in Table 2 and detailed conceptually in Section 3.4, we create:

- *Causal Augmentation Pairs* ($\mathcal{D}_{\text{causal}}$): Examples enforcing sensitivity to individual causal attributes C_j via **Attribute Upgradation** and **Degradation**, with standard preference labels ($\succ$).
- *Neutral Augmentation Pairs* ($\mathcal{D}_{\text{neutral}}$): Examples enforcing invariance to spurious attributes SP while ensuring C is irrelevant. These are generated via Irrelevant Pair Comparison (using query modification). These receive tie labels ($\approx$).

LLM prompts are in Appendix K. This yields the raw $\mathcal{D}_{\text{aug}}$.

3. Data Filtering. $\mathcal{D}_{\text{aug}}$ is filtered to $\mathcal{D}_{\text{aug\_filtered}}$ by retaining pairs where a baseline RM (trained on $\mathcal{D}_{\text{pref}}$) is uncertain or incorrect, focusing training on informative examples (details: Section 6, Appendix H.3). This yields the final training datasets $\mathcal{D}_{\text{pref}}$ and $\mathcal{D}_{\text{aug\_filtered}}$.

4.2. Robust Reward Model Training

Given the original data $\mathcal{D}_{\text{pref}}$ and the filtered augmented data $\mathcal{D}_{\text{aug\_filtered}}$, the final CROME reward model $\hat{\mathrm{R}}_\theta$ is trained by minimizing a composite loss function $\mathcal{L}(\theta)$ over the combined dataset $\mathcal{D} = \mathcal{D}_{\text{pref}} \cup \mathcal{D}_{\text{aug\_filtered}}$:

$$\mathcal{L}(\theta) = \underbrace{- \sum_{\substack{(\mathrm{Q},\mathrm{y}_w,\mathrm{y}_l) \\ \in \mathcal{D}_{\text{pref}} \cup \mathcal{D}_{\text{causal}}}} \log[\sigma(\Delta_{wl})]}_{\text{Preference Loss (Causal Sensitivity)}} \underbrace{- \lambda \sum_{\substack{(\mathrm{Q},\mathrm{A}_1,\mathrm{A}_2,\, y=\text{tie}) \\ \in \mathcal{D}_{\text{neutral}}}} \left(-\frac{1}{2} \left[\log \sigma(\Delta_{12}) + \log \sigma(-\Delta_{12}) \right] \right)}_{\text{Neutral Tie Loss (Spurious Invariance)}}$$

where $\Delta_{wl} = \hat{\mathrm{R}}_\theta(\mathrm{Q}, \mathrm{A}_w) - \hat{\mathrm{R}}_\theta(\mathrm{Q}, \mathrm{A}_l)$ and $\Delta_{12} = \hat{\mathrm{R}}_\theta(\mathrm{Q}, \mathrm{A}_1) - \hat{\mathrm{R}}_\theta(\mathrm{Q}, \mathrm{A}_2)$. The first term (Preference Loss) trains sensitivity to causal quality using $\mathcal{D}_{\text{pref}}$ and $\mathcal{D}_{\text{causal}}$. The second term (Neutral Tie Loss, weighted by $\lambda \geq 0$) trains invariance to spurious features using $\mathcal{D}_{\text{neutral}}$ by encouraging $\Delta_{12} \approx 0$ for tie-labeled pairs. For our current set of experiments we keep $\lambda = 1$.

This optimization guides $\hat{\mathrm{R}}_\theta$ to be sensitive to causal attributes C while remaining robust to variations in spurious attributes SP. We demonstrate CROME's effectiveness in mitigating reward hacking and improving downstream policy performance in Section 6.

5. Theoretical Analysis

We provide a theoretical analysis, detailed in Appendix I, to formalize how CROME's causal augmentation isolates true reward drivers from spurious correlates. Under an idealized model, we show that training on data with targeted interventions on causal attributes enables the learned reward model to accurately identify causal reward determinants, even in the presence of numerous, unspecified spurious features. To formalize this, we consider a setting where:

(i) Causal attributes $\mathrm{C}(\mathrm{A})$ and spurious attributes $\mathrm{SP}(\mathrm{A})$ are modeled as boolean variables.

(ii) True reward R^* is a sparse quadratic polynomial of $\mathrm{C}(\mathrm{A})$ only.

(iii) The learned $\hat{\mathrm{R}}_\theta$ can be a denser quadratic polynomial including $\mathrm{SP}(\mathrm{A})$ and $\mathrm{C}(\mathrm{A})\mathrm{SP}(\mathrm{A})$ terms.

(iv) Spurious attributes $\mathrm{SP}(\mathrm{A})$ are not descendants of causal attributes $\mathrm{C}(\mathrm{A})$.

(v) Causal augmentation is an ideal counterfactual that (given same exogenous factors leading to the answer) intervenes one $C_i \to \neg C_i$, leaving other C_j intervened to be their factual versions.

We frame learning the coefficients of R^* as an ℓ_1-constrained linear regression (Lasso) on features derived from attribute differences between an augmented answer A^{aug} and its original A. The key insight is that the feature matrix $\mathbf{F}$ from such augmented pairs exhibits properties conducive to sparse recovery, such as low column coherence or satisfying a Restricted Isometry Property (RIP) variant.

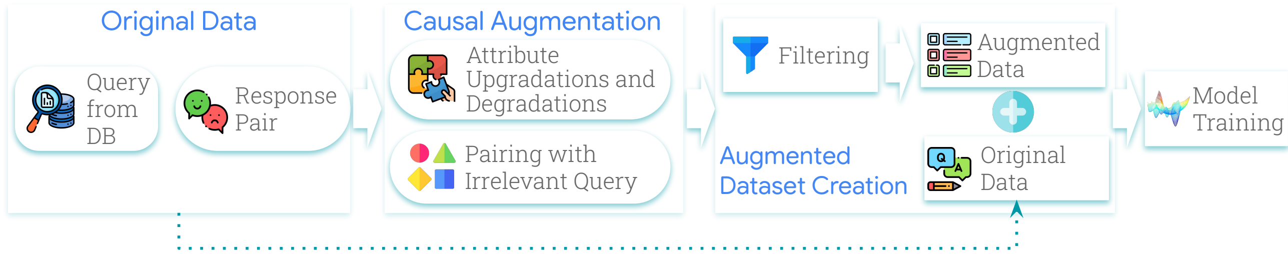


Figure 3: The CROME data augmentation pipeline. Original preference data ($\mathcal{D}_{\text{pref}}$) is used as a basis to generate: (1) *Causal Augmentations* ($\mathcal{D}_{\text{causal}}$) by performing **Attribute Upgradation and Degradation** on specific attributes to enforce sensitivity to genuine quality drivers, and (2) *Neutral Augmentations* ($\mathcal{D}_{\text{neutral}}$) via Irrelevant Query Neutrals (with tie-labels) to teach spurious feature invariance. After optional filtering, the reward model is trained on the combined original and augmented dataset.

Specifcally, compared to the original training set, the augmented one has a much lower RIP.

5.1. Main Theoretical Result (Informal)

This structure leads to the following result (formalized as Theorem I.3 in Appendix I):

Theorem 5.1 (Informal Statement). *Under the idealized model assumptions, ℓ_1-constrained regression on m causally augmented examples recovers the true causal reward coefficients* **a** *with an ℓ_2-error $\|\theta - \hat{\theta}\|_2$ that scales (ignoring constants and terms related to imperfect sparsity recovery) roughly as $O\left(\|\theta_{\mathcal{N}^c}\|_1(\frac{1}{k} + \sqrt{\frac{\log(k+\ell)}{m}})\right)$ where $\mathcal{N}$ is the top $O(k)$ coefficients in the R^* true reward model. This highlights a primary dependence on the number of causal attributes k and samples m, and only a weak, logarithmic dependence on the spurious attribute dimension ℓ.*

Implications: This theorem suggests that CROME's causal augmentation, by promoting favorable properties (like RIP or low incoherence) in the effective design matrix, guides the reward model towards genuine causal drivers. Further, the error vector has ℓ_2 norm is linear in the causal dimension k in the worst case and zero in the best case where R^* has sparser dependence on the causal factors. If it was the preference training dataset, the error could be proportional to $\|\theta\|_1$ (which is $O(k^2)$).

6. Experiments

Our experiments are designed to address the following research questions:

RQ1: RM Performance and Robustness: How does CROME perform on standard preference prediction tasks and how robust is it against spurious correlations(Table 3, Figure 4)?

RQ2: Best-of-N Alignment: Does the robustness achieved by CROME lead to favorable results in a Best-of-N setup as well, when compared to strong baselines (Figures 7, 8, Table 4)?

RQ3: Neutral Augmentations: How effective are different neutrals augmentation strategies in enforcing *invariance* to unknown spurious correlates (Figures 6, 9)?

6.1. Experimental Settings

CROME and baseline reward models (Vanilla RM, RRM (Liu et al., 2024)) are trained on the UltraFeedback dataset (Cui et al., 2023), with counterfactuals generated using Gemini 2.0 Flash. We evaluate performance on RewardBench (Lambert et al., 2024) and robustness on reWordBench (Wu et al., 2025) [2]. Experiments utilize diverse base LLMs (`Gemma-2-9B-IT`, `Qwen2.5-7B`, `Gemma-2-2B`) for both Pairwise Preference (PairPM) and Bradley-Terry (BT) reward models. Downstream alignment impact is assessed via Best-of-N selection on tasks including RewardBench, GSM8K, and WildGuardTest. Comprehensive details on datasets, model specifics, augmentation procedures, filtering, training hyperparameters, and all experimental configurations are provided in Appendix E.

6.2. Experimental Results addressing Research Questions (RQ1-3):

On **RewardBench** (Table 3), CROME consistently improves ranking accuracy over RRM across diverse base models and reward modeling techniques (PairPM, BT). These improvements are particularly notable on the challenging *Safety* (up to **13.18%**↑) and *Reasoning* (up to **7.19%**↑). CROME also demonstrates superior robustness on **reWordBench**, which tests for robustness of RMs against meaning-preserving transformations (Figure 4). With `Gemma-2-9B-IT`, CROME-PairPM shows an aggregate accuracy gain of up to **9.1%**↑ and is superior on (21/23) transformations.

[2] Since reWordBench has not been released, we follow the paper and communicated with the authors to reproduce it, see Appendix Section D

Table 3: Performance Comparison of Pairwise Preference Model and Bradley-Terry Reward Model on RewardBench trained using various base models

Base model	Method	PairPM Average	PairPM Chat	PairPM Chat-Hard	PairPM Safety	PairPM Reasoning	BT Average	BT Chat	BT Chat-Hard	BT Safety	BT Reasoning
Gemma-2-9B-IT	Vanilla RM	80.61	**98.18**	63.38	76.08	84.80	79.23	**97.49**	59.49	68.25	91.71
	RRM	82.53	96.93	72.04	73.78	87.36	83.22	97.49	**68.53**	73.18	93.68
	CROME	**87.93**	97.49	**72.70**	**86.96**	**94.55**	**85.33**	96.09	66.12	**82.84**	**96.27**
	$\Delta_{\text{CROME - RRM}}$	**+5.40↑**	**+0.56↑**	**+0.66↑**	**+13.18↑**	**+7.19↑**	**+2.11↑**	**-1.40↓**	**-2.41↓**	**+9.66↑**	**+2.59↑**
Qwen2.5-7B	Vanilla RM	78.18	**97.21**	52.85	73.99	88.68	72.73	97.21	46.27	68.04	79.39
	RRM	82.04	97.21	**64.80**	75.27	90.86	78.20	**98.04**	**59.65**	72.43	82.66
	CROME	**83.15**	96.37	61.73	**82.23**	**92.26**	**80.81**	96.93	58.66	**78.92**	**88.71**
	$\Delta_{\text{CROME - RRM}}$	**+1.11↑**	**-0.84↓**	**-3.07↓**	**+6.96↑**	**+1.40↑**	**+2.61↑**	**-1.11↓**	**-0.99↓**	**+6.49↑**	**+6.05↑**
Gemma-2-2B	Vanilla RM	53.75	92.88	33.33	42.03	46.74	65.52	94.27	38.27	50.20	79.34
	RRM	66.23	**94.13**	43.75	47.64	79.38	66.95	**94.97**	49.34	50.07	73.42
	CROME	**70.69**	92.18	**50.00**	**55.14**	**85.42**	**72.45**	92.74	**53.62**	**60.00**	**83.45**
	$\Delta_{\text{CROME - RRM}}$	**+4.46↑**	**-1.95↓**	**+6.25↑**	**+7.50↑**	**+6.04↑**	**+5.50↑**	**-2.23↓**	**+4.28↑**	**+9.93↑**	**+10.03↑**

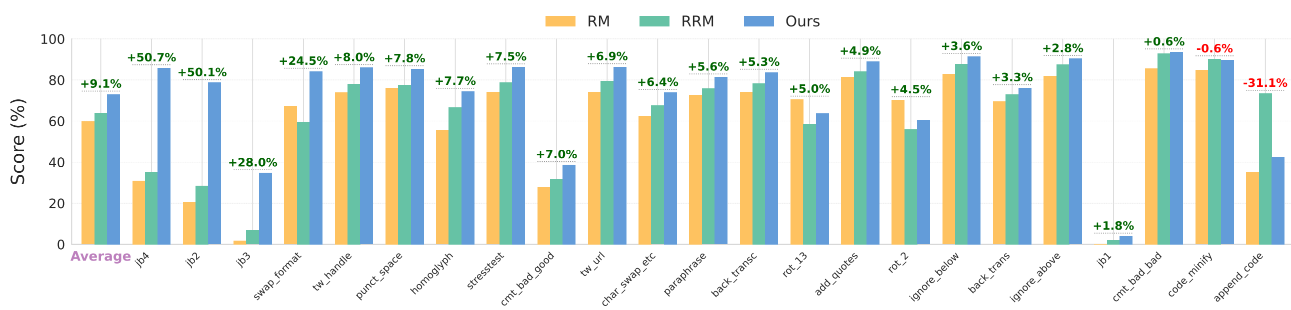


Figure 4: Absolute robustness comparison of RM, RRM and CROME on reWordBench.

Key Takeaway: CROME improves RM performance on standard benchmarks while significantly improving performance and mitigating ranking accuracy drops on diverse transformed inputs, *without ever being explicitly trained on such spurious transformations*.

N	CROME vs RM: CROME	CROME vs RM: RM	CROME vs RM: Ties	CROME vs RRM: CROME	CROME vs RRM: RRM	CROME vs RRM: Ties
4	**28.08**	13.85	58.07	**28.03**	14.13	57.84
8	**34.32**	17.24	48.43	**34.36**	17.19	48.45
16	**39.93**	20.54	39.53	**41.14**	20.40	38.46
32	**44.79**	21.88	33.33	**45.46**	22.01	32.53

Table 4: Win rates for CROME, on RewardBench

BoN for Robust LLM Alignment Across Chat, Reasoning, and Safety Following the method used by Wu et al. (2025), in Table 4 we perform best-of-n selection using CROME across RewardBench categories, which consists of datasets such as AlpacaEval. Across all values of N, CROME provided significant improvements over baselines in a head-to-head comparison.

Key Takeaway: CROME's emphasis on causal attributes enhances its discriminative power in Best-of-N selection, leading to more consistent identification of superior responses.

Ranking Accuracy Improvements: In Fig. 5, we test difference between RewardBench and reWordBench scores (following the macro-avg metric used in Wu et al. (2025)). CROME exhibits a smaller ranking accuracy drop from RewardBench to reWordBench (In case of PairPM: 19.78% vs. RRM's 21.54%. See Appendix C and D for more details.

Key Takeaway: Assuming sufficient concentration of spurious elements in the prompt as well as the N re-

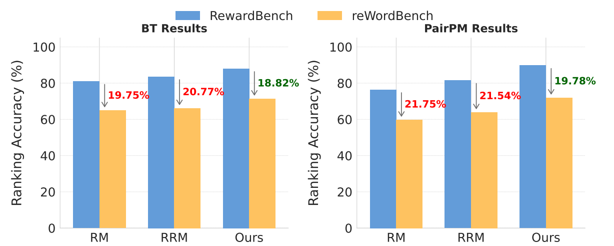


Figure 5: % improvement in ranking accuracy between RewardBench and reWordBench.

sponses, CROME is better at selecting the best response based on causal attributes only. For e.g., in safety, harmful prompts and responses may be disguised as benign.

Causal Attributes help in detecting jailbreaks In Fig. 7, BoN with CROME shows significant improvements on safety as measured on WildGuardTest (`Gemma-2-9B-IT` base model). In particular the attack success ratio (ASR) on harmful prompts is much lower compared to models aligned with RM and RRM and this gap increases with N. This improved ASR comes at at a similar refusal-to-answer rate on benign prompts.

Key Takeaway: CROME's causal augmentations achieve a superior trade-off between safety and over-refusals, because its contrastive pairs delineate the decision boundary for harmful content more faithfully. This leads to safer content, while avoiding excessive refusals on benign prompts.

Disentangling Content related features from stylistic (spurious) ones helps in reasoning In Fig. 8, CROME shows a consistent gap over baseslines across different values of N on GSM8K (`Gemma-2-9B-IT` base model). Non robust reward models may focus on stylistic details. Good looking, detailed but wrong reasoning steps may misguide non-robust RMs into giving a higher score to the response.

Key Takeaway: Reasoning correctness is dependent on focusing on correctness over stylistic features. Our training ensures CROME is good at capturing content-features over other attributes.

Neutrals help in spurious suppression Neutral augmentations significantly improve robustness compared to causal-only training (Figures 6 and 9). All neutral variants outperform the causal-only CROME-C model. Among them, CROME-IQN achieves the best overall performance on RewardBench, with a gain of **+5.4%**↑ over the RRM baseline. Meanwhile, CROME-CAN achieves the best performance on reWordBench, with a gain of **+12.5%**↑.

Key Takeaway: *Explicit* suppression of spurious correlates via neutral augmentations mitigates reward hacking by learning *invariant* reward signals, thereby improving downstream performance.

In Fig. 6, we plot the performance on subsets of RewardBench, of different CROME variants including: CROME-C (only causals), CROME-IQN (causals + irrelevant query neutrals), CROME-PARA (causals + paraphrased neutrals), CROME-CAN (causals + causally-aligned neutrals), and CROME-IQN+CAN (causals + irrelevant query neutrals + causally-aligned neutrals). On the especially challenging *Chat-Hard* subset, CROME-IQN performs best. We discuss the reasons for these variations in Appendix F. Prompts for obtaining these neutrals is given in Appendix K.

Key Takeaway: A combination of well-designed augmentation strategies, e.g. causal upgradations and degradations, along with irrelevant query neutrals, produces the most robust and generalizable reward models.

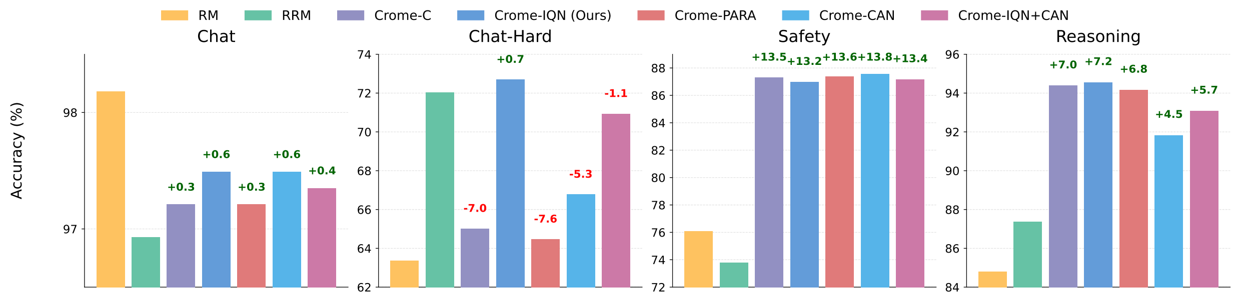


Figure 6: Evaluations of neutral augmentation variants on the different subsets of RewardBench.

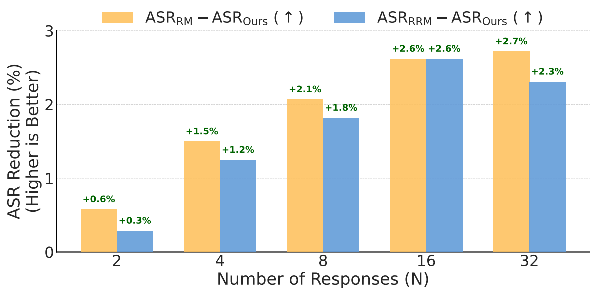


Figure 7: Best-of-N results: ASR reduction on WildGuardTest.

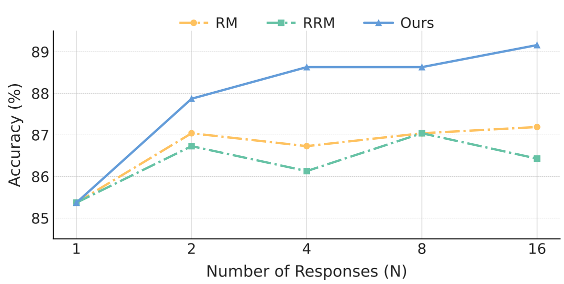


Figure 8: Best-of-N Reasoning evaluation on GSM8K.

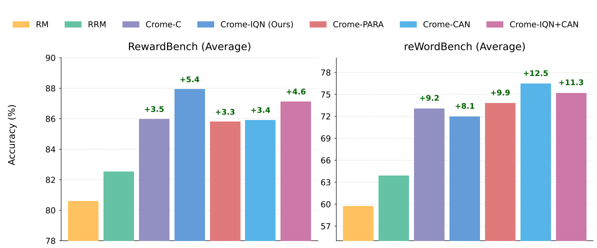


Figure 9: Average performance on RewardBench and reWordBench for CROME trained with different neutral augmentation strategies.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

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

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

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Chen, L., Zhu, C., Soselia, D., Chen, J., Zhou, T., Goldstein, T., Huang, H., Shoeybi, M., and Catanzaro, B. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024.

Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., Liu, T., and Han, B. Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37:96640–96670, 2024.

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

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Coste, T., Anwar, U., Kirk, R., and Krueger, D. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

Eisenstein, J., Nagpal, C., Agarwal, A., Beirami, A., D'Amour, A., Dvijotham, D., Fisch, A., Heller, K., Pfohl, S., Ramachandran, D., et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Feder, A., Oved, N., Shalit, U., and Reichart, R. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Gupta, T., Shandilya, S., Zhang, X., Madhavan, R., Ghosh, S., Bansal, C., Yao, H., and Rajmohan, S. Carmo: Dynamic criteria generation for context-aware reward modelling, 2025. URL https://arxiv.org/abs/2410.21545.

Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., and Dziri, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024. URL https://arxiv.org/abs/2406.18495.

Hong, J., Lee, N., and Thorne, J. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Kaushik, D., Hovy, E., and Lipton, Z. C. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Khalifa, M., Agarwal, R., Logeswaran, L., Kim, J., Peng, H., Lee, M., Lee, H., and Wang, L. Process reward models that think. *arXiv preprint arXiv:2504.16828*, 2025.

Kiciman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Li, W. and Li, Y. Process reward model with q-value rankings. *arXiv preprint arXiv:2410.11287*, 2024.

Liu, T., Xiong, W., Ren, J., Chen, L., Wu, J., Joshi, R., Gao, Y., Shen, J., Qin, Z., Yu, T., et al. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*, 2024.

Liu, Y., Yao, Z., Min, R., Cao, Y., Hou, L., and Li, J. Pairwise rm: Perform best-of-n sampling with knockout tournament. *arXiv preprint arXiv:2501.13007*, 2025.

Long, S., Piché, A., Zantedeschi, V., Schuster, T., and Drouin, A. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lou, X., Yan, D., Shen, W., Yan, Y., Xie, J., and Zhang, J. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*, 2024.

Meng, Y., Xia, M., and Chen, D. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Mishra, A., Nayak, G., Bhattacharya, S., Kumar, T., Shah, A., and Foltin, M. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1538–1545, 2024.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. 2012.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pace, A., Mallinson, J., Malmi, E., Krause, S., and Severyn, A. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086*, 2024.

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models, 2022. URL https://arxiv.org/abs/2201.03544.

Park, R., Rafailov, R., Ermon, S., and Finn, C. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Qiang, Y., Nandi, S., Mehrabi, N., Steeg, G. V., Kumar, A., Rumshisky, A., and Galstyan, A. Prompt perturbation consistency learning for robust language models. *arXiv preprint arXiv:2402.15833*, 2024.

Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., Shen, J., Liu, T., Liu, J., Metzler, D., et al. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Ramé, A., Vieillard, N., Hussenot, L., Dadashi, R., Cideron, G., Bachem, O., and Ferret, J. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*, 2024.

Ravfogel, S., Svete, A., Snæbjarnarson, V., and Cotterell, R. Gumbel counterfactual generation from language models, 2025. URL https://arxiv.org/abs/2411.07180.

Reber, D., Richardson, S., Nief, T., Garbacea, C., and Veitch, V. Rate: Score reward models with imperfect rewrites of rewrites. *arXiv preprint arXiv:2410.11348*, 2024.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

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

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shen, J., Xu, R., Jun, Y., Qin, Z., Liu, T., Yang, C., Liang, Y., Baumgartner, S., and Bendersky, M. Boosting reward model with preference-conditional multi-aspect synthetic data generation. *arXiv preprint arXiv:2407.16008*, 2024.

Shen, L., Chen, S., Song, L., Jin, L., Peng, B., Mi, H., Khashabi, D., and Yu, D. The trickle-down impact of reward (in-) consistency on rlhf. *arXiv preprint arXiv:2309.16155*, 2023.

Singhal, P., Goyal, T., Xu, J., and Durrett, G. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

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

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Tu, R., Ma, C., and Zhang, C. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Wang, C., Zhao, Z., Jiang, Y., Chen, Z., Zhu, C., Chen, Y., Liu, J., Zhang, L., Fan, X., Ma, H., et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.

Wang, H., Xiong, W., Xie, T., Zhao, H., and Zhang, T. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024.

Wu, Z., Yasunaga, M., Cohen, A., Kim, Y., Celikyilmaz, A., and Ghazvininejad, M. rewordbench: Benchmarking and improving the robustness of reward models with transformed inputs. *arXiv preprint arXiv:2503.11751*, 2025.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Yang, R., Pan, X., Luo, F., Qiu, S., Zhong, H., Yu, D., and Chen, J. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024b.

Zhang, L., Hosseini, A., Bansal, H., Kazemi, M., Kumar, A., and Agarwal, R. Generative verifiers: Reward modeling as next-token prediction. 2025.

Zhang, X., Xiong, W., Chen, L., Zhou, T., Huang, H., and Zhang, T. From lists to emojis: How format bias affects model alignment. *arXiv preprint arXiv:2409.11704*, 2024.

Zhao, J., Liu, R., Zhang, K., Zhou, Z., Gao, J., Li, D., Lyu, J., Qian, Z., Qi, B., Li, X., et al. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*, 2025.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Zhu, X., Tan, C., Chen, P., Sennrich, R., Zhang, Y., and Hu, H. Charm: Calibrating reward models with chatbot arena scores. *arXiv preprint arXiv:2504.10045*, 2025.

Supplementary Material

These supplementary materials provide additional details, derivations, and experimental results for our paper. The appendix is organized as follows:

- Section A discusses potential limitations of this work.
- Section B provides a broader overview of recent related literature. This is an expanded version of the literature covered in the main paper.
- Section C provides some additional set of results. This is an expanded version of the results covered in the main paper.
- Section D provides the detailed steps we took to reproduce the reWordBench benchmark, as proposed in Wu et al. (2025).
- Section E provides a detailed overview of our experimental setup.
- Section F provides a detailed walk through of how our causal model extends to prior method. We revisit prior works in light of our causal model. It extends on the shorter version provided in Section 3.
- Section G provides a walkthrough of the causal details of the core data augmentation strategies.
- Section H provides a detailed walk through of the method used to train the reward model. It extends on the shorter version provided in Section 4.
- Section I provides a detailed analysis of the theory relating to Reward Hacking and how our proposed method mitigates it.
- Section J presents a qualitative example of augmented data created from original data using which is used to train CROME.
- Section K presents a lists of prompt templates that we use to query our models for generating the data.
- Section L presents a qualitative view common failure modes or biases commonly observed in reward models.

A. Limitations and Future Work

While CROME demonstrates significant improvements, we acknowledge certain limitations which also suggest avenues for future research:

- **Idealized Assumptions in Theoretical Analysis:** Our theoretical justification (Section 5, Appendix I) relies on simplifying assumptions such as boolean attributes, quadratic reward models, and perfect counterfactual interventions. These idealizations, necessary for analytical tractability, mean our formal guarantees are indicative of CROME's potential mechanism rather than absolute predictions of real-world performance, where the complexities of LLM behavior and data are greater.
- **Scalability and Cost of Data Augmentation:** The generation of targeted causal and neutral augmentations, while effective, involves multiple LLM inference calls per original data point. Although filtering helps optimize the final dataset size, the initial augmentation phase can be computationally intensive and potentially costly for extremely large-scale applications. Future work could explore more sample-efficient augmentation strategies or methods to distill the benefits of augmentation into smaller datasets.
- **Generalization to Highly Novel Spurious Correlations:** CROME is designed to be robust against unspecified spurious correlations by focusing on causal signals and diverse neutral examples. However, its ability to generalize to entirely novel types of spuriousness, drastically different from any patterns implicitly covered or contrasted during augmentation, remains an empirical question. The breadth and nature of the neutral augmentations play a role here, and continuous adaptation or more abstract invariance learning might be needed for extreme out-of-distribution spuriousness.

- **Fidelity of LLM-Generated Counterfactuals:** The efficacy of CROME is linked to the quality of the LLM-generated counterfactuals. While current LLMs are powerful, ensuring perfect attribute isolation in causal augmentations or complete causal content preservation in neutral pairs is challenging. Imperfections in these LLM-approximated interventions can introduce noise. While our empirical results show strong benefits, further research into enhancing the precision and verifiability of LLM-driven textual counterfactual generation could yield additional improvements.

Future research could focus on extending the theoretical framework to encompass more realistic settings, developing more cost-effective and adaptive augmentation techniques, and further exploring the boundaries of generalization against emergent spurious correlations.

B. Extended Related Works

Our work on CROME, a framework for causally robust reward modeling, intersects with and builds upon several key areas of research: the alignment of Large Language Models (LLMs) via human feedback, techniques for reward model training, the persistent challenge of reward hacking, the application of causal inference principles to machine learning, and data augmentation strategies for enhancing model robustness.

LLM Alignment and RLHF. The dominant paradigm for steering LLM behavior towards desired attributes like helpfulness, honesty, and harmlessness is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a; Askell et al., 2021). The standard RLHF process involves training a reward model (RM) on human preferences (typically pairwise comparisons) and subsequently using this RM as a reward signal to fine-tune the LLM policy via RL algorithms such as PPO (Schulman et al., 2017). The quality, calibration, and robustness of the RM are paramount, as flaws in the RM directly impact the alignment outcome (Casper et al., 2023). While alternative alignment algorithms like Direct Preference Optimization (DPO) (Rafailov et al., 2024) and its extensions (e.g., IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), ORPO (Hong et al., 2024), SimPO (Meng et al., 2024)) bypass explicit RM training by directly optimizing the policy on preference data, they still implicitly rely on the preference information learnable from the data, making the problem of distinguishing true quality from spurious correlates equally relevant.

Reward Modeling Techniques. Learning accurate reward models from preference data remains a central challenge. Methodologies include Bradley-Terry style pointwise models that learn a scalar score $r(x, y)$ (Bradley & Terry, 1952; Ouyang et al., 2022; Bai et al., 2022a), and pairwise ranking models that directly predict preference probabilities, often implemented within the LLM architecture itself (PairPM) (Liu et al., 2025; Qin et al., 2023). Other approaches explore Q-function based rewards (Li & Li, 2024) or process supervision (Khalifa et al., 2025). Significant effort focuses on improving specific RM properties like calibration (Zhu et al., 2025; Zhao et al., 2023), training efficiency (Tunstall et al., 2023), uncertainty quantification (Lou et al., 2024), interpretability through multi-aspect rewards (Wang et al., 2024; Yang et al., 2024b), and scalability via reasoning or chain-of-thought mechanisms (Zhang et al., 2025; Zhao et al., 2025). Our work complements these efforts by focusing specifically on enhancing the causal **robustness** of the learned reward function $\hat{R}$ against spurious attributes.

Reward Hacking and Spurious Correlations. Learned reward models are notoriously susceptible to *reward hacking* or *over-optimization* (Gao et al., 2023; Skalse et al., 2022; Pan et al., 2022). Because RMs are trained on finite, potentially biased data, they often learn to associate high rewards with superficial or *spurious* features that are merely correlated with desirable responses in the training set. Common examples include excessive length or verbosity (Singhal et al., 2023), specific formatting patterns like lists or markdown (Zhang et al., 2024), adherence to stylistic conventions like politeness, or even sycophantic agreement with user views (Denison et al., 2024). Policies optimized against such RMs learn to exploit these spurious cues, leading to outputs that maximize the predicted reward but fail to align with genuine human preferences or task goals (Shen et al., 2023).

Approaches to Mitigating Reward Hacking. Various strategies have been proposed to address reward hacking. Model-centric approaches include using ensembles of RMs to average out idiosyncratic biases (Coste et al., 2023; Eisenstein et al., 2023; Ramé et al., 2024), incorporating explicit calibration methods (Zhao et al., 2023), or designing architectures that factorize reward components, such as ODIN's disentanglement of quality and length (Chen et al., 2024). Policy-optimization techniques might involve adding explicit penalties for spurious features (e.g., length penalties (Park et al., 2024)) or using specific regularization methods during fine-tuning. Data-centric approaches aim to improve the training data or process itself. Examples include iterative re-labeling or refinement (Bai et al., 2022b), performing consistency checks across related prompts (Shen et al., 2023), or augmenting the dataset with synthetic examples designed to improve robustness (Pace et al., 2024; Shen et al., 2024). Our work, CROME, falls firmly in this data-centric category. It is closely related to RRM (Liu et al., 2024), which also uses data augmentation (non-contextual and neutrals) for robustness. However, CROME is distinct in its use of an explicit causal framework and its generation of targeted, attribute-specific counterfactuals to disentangle causal from spurious factors.

Causal Inference in Machine Learning. Causal inference provides formal tools, such as Structural Causal Models (SCMs) and DAGs (Pearl, 2009; Peters et al., 2017), for reasoning about cause-effect relationships, confounding, and counterfactuals. Applying causal principles in machine learning aims to build models that are more robust, fair, and interpretable by focusing on underlying causal mechanisms rather than potentially brittle statistical correlations (Schölkopf et al., 2021). Techniques like Invariant Risk Minimization (IRM) seek models that perform well across different environments by relying on invariant (presumably causal) predictors (Arjovsky et al., 2019). Our work adopts this causal perspective, framing spurious attributes as non-causal factors whose influence on the learned reward model should be minimized.

Causality in LLMs and NLP. The intersection of causality and LLMs is rapidly evolving. Research includes probing the innate causal reasoning abilities of LLMs (Kiciman et al., 2023; Chi et al., 2024), leveraging LLMs as tools for automating parts of the causal discovery or analysis pipeline (Long et al., 2023; Tu et al., 2023), and applying causal methods to enhance NLP tasks. For instance, counterfactual reasoning and data augmentation have been used to improve robustness against biases in text classification (Kaushik et al., 2019; Feder et al., 2021) and assess fairness (Feder et al., 2022). CROME uniquely employs a predefined causal graph to structure the generation of counterfactual data specifically for training a robust RM, using LLMs as the generation engine.

Data Augmentation for Robustness. Data augmentation is a cornerstone technique for improving model generalization. Beyond traditional NLP methods like synonym replacement or back-translation (Wu et al., 2025), more recent approaches leverage LLMs for sophisticated augmentations, including paraphrasing, style transfer, generating adversarial examples (Qiang et al., 2024), or creating counterfactuals (Mishra et al., 2024; Feder et al., 2021). Counterfactual generation, often using LLMs as rewriters, is also central to evaluation methods like RATE (Reber et al., 2024), which uses "rewrites of rewrites" to estimate causal effects robustly. Methods based on sampling, like Gumbel temperature sampling, have also been explored for counterfactual generation (Ravfogel et al., 2025). In the specific context of reward modeling, data augmentation aims to enhance robustness against spurious correlations; examples include the non-contextual and query-independent pairs used by RRM (Liu et al., 2024) or consistency checks via paraphrased inputs as explored in REWORDBENCH (Wu et al., 2025). Furthermore, generating entirely synthetic preference pairs (Pace et al., 2024; Shen et al., 2024) represents another data-centric approach to improving reward models. Counterfactual data augmentation, particularly generating minimally different pairs to isolate specific features (Kaushik et al., 2019), is highly relevant to disentangling causal factors. Our work, CROME, operationalizes this concept within an explicit causal framework, generating targeted "causal" (attribute-isolating) and "neutral" (spurious-varying) pairs via LLM rewriting to enforce specific invariance and sensitivity properties in the trained RM.

Positioning of CROME. CROME integrates insights from causal inference and data augmentation to address the critical problem of reward hacking in LLM alignment. While related works like RRM (Liu et al., 2024) use data augmentation for robustness and CARMO (Gupta et al., 2025) uses LLMs for criteria generation, CROME is distinguished by its explicit grounding in a causal graph model of answer attributes. It systematically generates attribute-specific counterfactual and neutral examples via guided LLM prompting to directly train the RM to distinguish causal quality drivers (C) from spurious correlates (SP). This allows CROME to potentially handle a wider range of spurious attributes beyond commonly studied ones like length, aiming for a more principled and generalizable form of robustness. We provide the methodology and empirical validation (Section 6) demonstrating that this causally-informed data augmentation leads to more robust reward models and better downstream policy alignment compared to standard baselines.

C. Additional Results

Our main findings presented in this section are as follows:

- **Stable and Significant Performance Gains:** CROME consistently outperforms baseline reward models (Vanilla RM and RRM) on RewardBench across multiple independent training runs, with small standard deviations indicating stable performance. The improvements, particularly on reWordBench transformations, are substantial and typically exceed multiple standard deviations of the baselines, underscoring their statistical significance (Sec. C.1, C.2).
- **Robustness to Oracle LLM Choice:** CROME's methodology for attribute extraction and augmentation demonstrates effectiveness even when using different oracle LLMs (e.g., transitioning from `Gemini-2.0-Flash` to `Gemma-3-27B-IT`), achieving notable improvements in robustness, such as an average of +4.0% on reWordBench with `Gemma-3-27B-IT` (Sec. C.3).
- **Strong Out-of-Distribution Generalization:** CROME exhibits strong generalization from in-distribution (UltraFeedback validation) to out-of-distribution benchmarks (RewardBench, reWordBench). Notably, it often achieves the highest OOD accuracy (e.g., +7.02% over RRM on reWordBench PairPM) while having similar ID accuracy, suggesting its augmentations teach more generalizable preference representations (Sec. C.4).

C.1. Variance in Performance on RewardBench

To assess the stability of our findings, we conducted three independent training runs for reward models built upon the `Gemma-2-9B-IT` base model. Table 5 for PairPM and BT reports the mean accuracy and standard deviation on **RewardBench** categories. The standard deviations for average RewardBench accuracies are consistently small across all methods (e.g., ± 0.09 on average for CROME-PairPM, ± 0.12 on average for RRM-PairPM), indicating stable performance. While there is some variation in specific sub-categories, CROME's average performance advantage over baselines remains robust.

Table 5: Mean Accuracy and Standard Deviation across 3 different training runs of `Gemma-2-9B-IT` based Reward Models in both PairPM and Bradley-Terry Reward Model settings. Results on RewardBench.

	Method	PairPM					BT				
		Average	Chat	Chat-Hard	Safety	Reasoning	Average	Chat	Chat-Hard	Safety	Reasoning
Gemma-2-9B-IT	Vanilla RM	81.22 ± 0.56	**97.90 ± 0.48**	63.64 ± 0.28	77.48 ± 1.21	85.88 ± 1.34	79.14 ± 0.68	**97.26 ± 0.40**	58.85 ± 1.14	69.30 ± 3.61	91.17 ± 1.17
	RRM	82.54 ± 0.12	97.12 ± 0.21	71.05 ± 0.87	74.70 ± 0.98	87.27 ± 0.21	83.46 ± 0.26	97.21 ± 0.28	**69.15 ± 0.54**	73.13 ± 0.61	94.35 ± 0.59
	CROME	**87.84 ± 0.09**	97.54 ± 0.21	**72.30 ± 0.39**	**87.14 ± 0.16**	**94.39 ± 0.21**	**85.46 ± 0.27**	96.28 ± 0.32	65.83 ± 0.81	**84.05 ± 1.10**	**95.70 ± 0.52**
	$\Delta_{\text{CROME - RRM}}$	**+5.30↑**	**+0.42↑**	**+1.25↑**	**+12.44↑**	**+7.12↑**	**+2.00↑**	**-0.93↓**	**-3.32↓**	**+10.92↑**	**+1.35↑**

Remark C.1. Note that main paper Table 3 have results of only the first out of the three training runs considered in these variance experiments.

C.2. Variance in Performance on reWordBench

For **reWordBench**, we plot mean performance numbers and error bars showing std. deviation in Figures 10 and 11. Here we depict mean accuracies with error bars representing standard deviations. Across most transformations, the error bars are relatively small, particularly for the average performance over all transformations. The observed improvements of CROME compared to RRM and Vanilla RM are substantial and typically exceed multiple standard deviations of the respective models, suggesting that these gains are statistically significant.

C.3. Experiments with a different Oracle LLM: `Gemma-3-27B-IT`

CROME depends on an Oracle LLM for generating attributes and creating augmented data. In the main paper we use a low cost API based model, `Gemini-2.0-Flash`. In this section we show that our method is effective at improving robustness, even with open-weights based models. In particular, we use `Gemma-3-27B-IT` to perform attribute extraction and augmentations following which we train CROME on the augmented data. See Table 6 for RewardBench results and Figure 12 for reWordBench results. Our results indicate an average improvement of 4.0% with an improvement in 18/23 transformations.

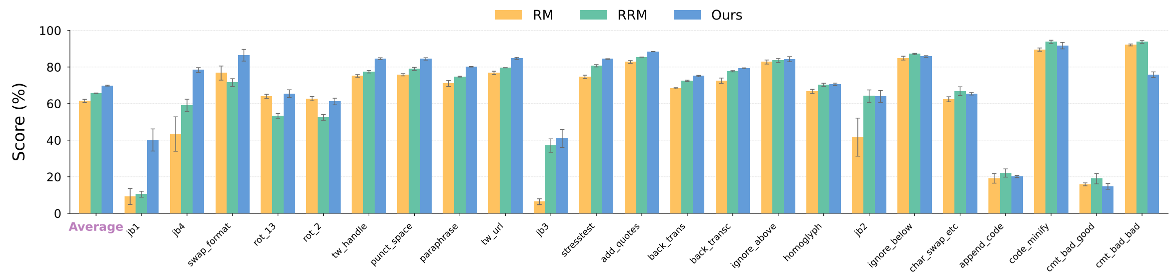


Figure 10: **Standard deviation error-bars** for absolute robustness comparison of RM, RRM and CROME in the **Bradley-Terry setup**, for reward models built over `Gemma-2-9B-IT`. Mean values and std deviation plotted are for 3 independent training runs.

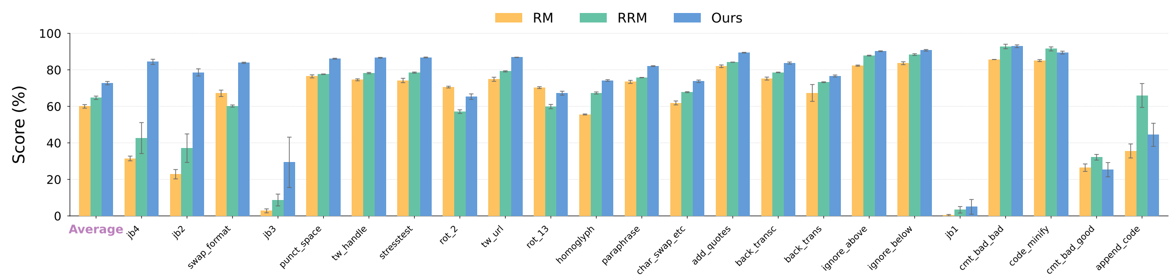


Figure 11: **Standard deviation error-bars** for absolute robustness comparison of RM, RRM and CROME in the **PairPM setup**, for reward models built over `Gemma-2-9B-IT`. Mean values and std deviation plotted are for 3 independent training runs.

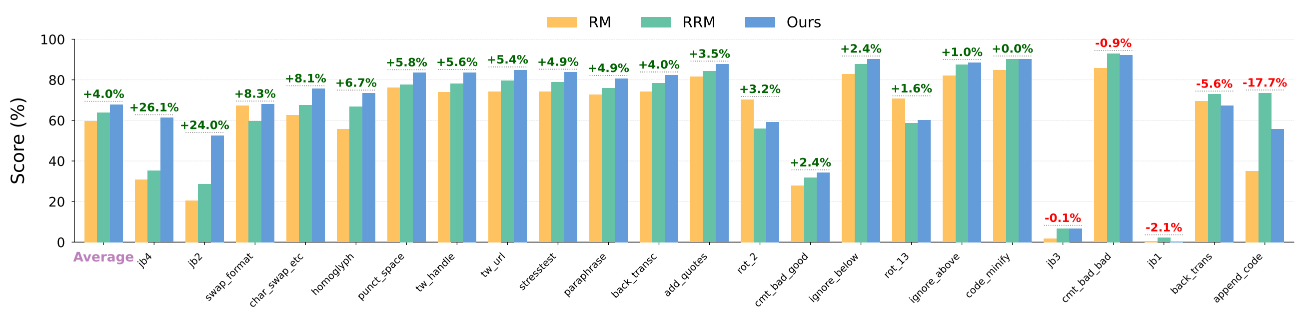


Figure 12: Absolute robustness comparison of RM, RRM and CROME on reWordBench. Here all reward models are `Gemma-2-9B-IT` based, in the PairPM setting.

C.4. Effective Robustness of CROME and Baselines

We evaluate the generalization capabilities of the trained reward models by comparing their performance on in-distribution (ID) data (UltraFeedback validation split) against out-of-distribution (OOD) benchmarks (RewardBench, reWordBench). Table 7 presents these results for models based on `Gemma-2-9B-IT`. CROME demonstrates strong OOD performance, particularly on reWordBench. For instance, in the PairPM setup, CROME achieves the highest reWordBench accuracy (72.71%), while having similar ID accuracy, suggesting that its learned robustness translates well to challenging, unseen transformations. Similarly, for Bradley Terry models, CROME shows the best reWordBench accuracy (69.81%) and similar ID accuracies compared to baselines. Overall, these results indicate that CROME's augmentations effectively teach more generalizable representations of preferences.

Table 6: Performance of Pairwise Preference Methods (PairPM) on RewardBench with `Gemma-2-9B-IT` as base model and `Gemma-3-27B-IT` as oracle LLM used for augmentations.

Method	PairPM				
	Average	Chat	Chat-Hard	Safety	Reasoning
Vanilla RM	80.61	**98.18**	63.38	76.08	84.80
RRM	82.53	96.93	72.04	73.78	87.36
CROME (`Gemma-3-27B-IT` Oracle)	85.15	97.21	68.75	83.51	91.13
CROME (`Gemini-2.0-Flash` Oracle)	**87.93**	97.49	**72.70**	**86.96**	**94.55**

Table 7: Comparison of In-Distribution (UltraFeedback-Val) and Out-of-Distribution (RewardBench, reWordBench) Accuracy (%) for `Gemma-2-9B-IT` RMs

PairPM							
Model	Ultrafeedback (ID)	reWordBench Accuracy (OOD)	RewardBench Accuracy (OOD)				
			Chat	Chat-Hard	Safety	Reasoning	Avg
RM	74.55	59.97	**97.90**	63.64	77.48	85.88	81.22
RRM	**75.20**	64.68	97.12	71.05	74.70	87.27	82.54
Ours	74.02	**72.71**	97.54	**72.30**	**87.14**	**94.39**	**87.84**
Bradley Terry							
Model	Ultrafeedback (ID)	reWordBench Accuracy (OOD)	RewardBench Accuracy (OOD)				
			Chat	Chat-Hard	Safety	Reasoning	Avg
RM	74.60	61.48	**97.26**	58.85	69.30	91.17	79.14
RRM	**74.75**	65.69	97.21	**69.15**	73.13	94.35	83.46
Ours	74.00	**69.81**	96.28	65.83	**84.05**	**95.70**	**85.46**

C.5. Extended Results on Safety Prompts from WildGuardTest

To complement the Best-of-N (BoN) safety results in Figure 7 (Sec. 6.2), we provide the complete Attack Success Rate (ASR) on harmful prompts and Refusal to Answer (RTA) on benign prompts in Table 8. We note that lower numbers are better for both ASR as well as RTA. Significantly, the results indicate that without too much regression on RTA ($< 0.5\%$ decrease), we show consistent gains in ASR (%) numbers and these gains increase as N becomes larger. For instance, at N=32, CROME reduces ASR to **39.39%**, compared to 42.11% for RM and 41.70% for RRM. In practice, reward models are used to detect jailbreak attacks, and hence our model performance indicates a favorable trade-off as the reward model detects harmful content (resisting jail-break attempts) while maintaining utility (low refusal-to-answer rate).

C.6. Additional Results on reWordBench

We provide additional results on reWordBench in this section. See Figures 13 to 17 for reWordBench results on various base models over which we build our Reward Models, such as `Gemma-2-9B-IT`, `Gemma-2-2B` and `Qwen2.5-7B`, across Bradley-Terry and pairwise-preference Reward Models.

D. reWordBench Reproduction

The primary motivation reWordBench is the observation that contemporary reward models—key components of RLHF systems—often latch onto superficial formatting cues or benign artifacts in their training data, leading to dramatic drops in pairwise-preference accuracy under minor, semantically neutral edits. To diagnose and quantify this brittleness in a systematic

Table 8: Comparison of Attack Success Rate (ASR) on harmful prompts and Refusal to Answer (RTA) on benign prompts for CROME compared to baselines (RM, RRM) in the Best-of-N setup for varying N. Lower values are considered better for both metrics.

	RM		RRM		**Ours**	
N	ASR (%)	RTA (%)	ASR (%)	RTA (%)	ASR (%)	RTA (%)
2	32.76	**7.39**	32.47	**7.39**	**32.18**	7.58
4	36.13	**6.97**	35.88	7.18	**34.63**	7.46
8	38.49	6.29	38.24	**6.10**	**36.42**	6.97
16	39.33	6.27	39.33	**5.89**	**36.71**	6.39
32	42.11	**5.80**	41.70	6.30	**39.39**	6.01

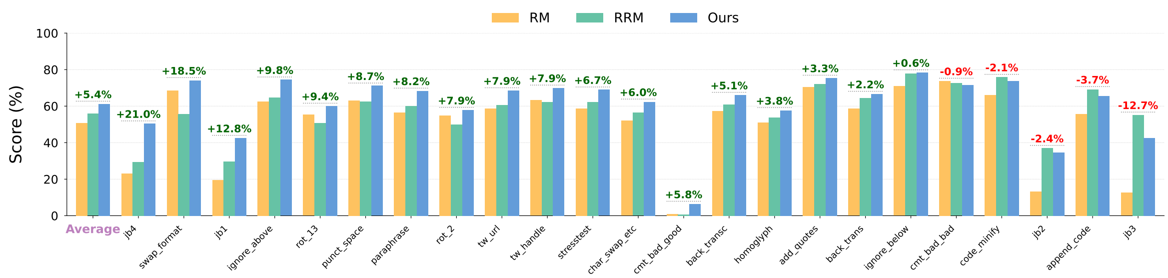


Figure 13: Absolute Robustness Comparison of RM, RRM and CROME in the Bradley-Terry RM setup, for reward models built over `Gemma-2-2B-IT`.

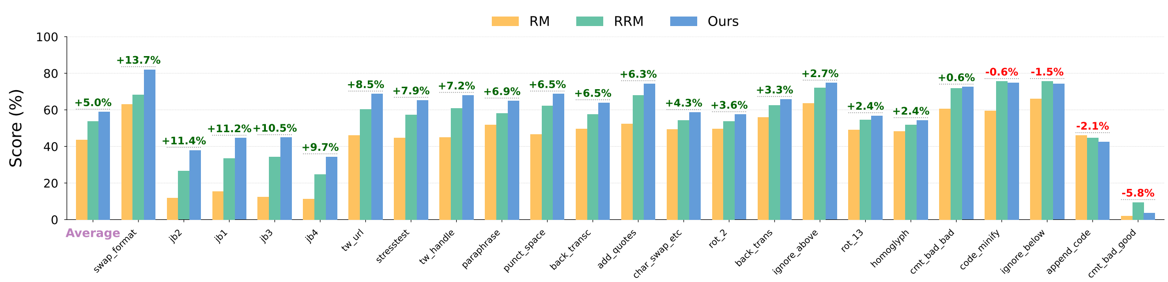


Figure 14: Absolute Robustness Comparison of RM, RRM and CROME in the PairPM setup, for reward models built over `Gemma-2-2B-IT`.

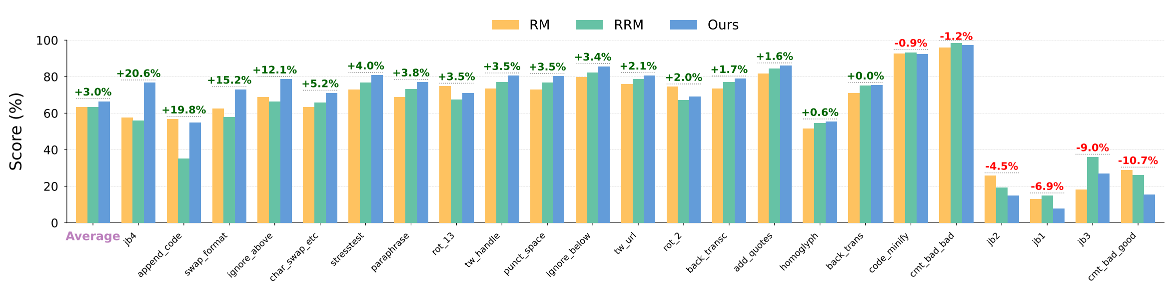


Figure 15: Absolute Robustness Comparison of RM, RRM and CROME in the PairPM setup, for reward models built over `Qwen2.5-7B`.

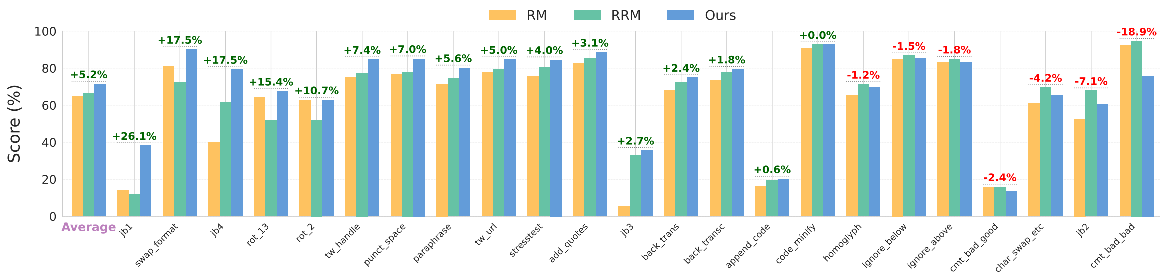


Figure 16: Absolute Robustness Comparison of RM, RRM and CROME in the Bradley-Terry RM setup, for reward models built over `Gemma-2-9B-IT`.

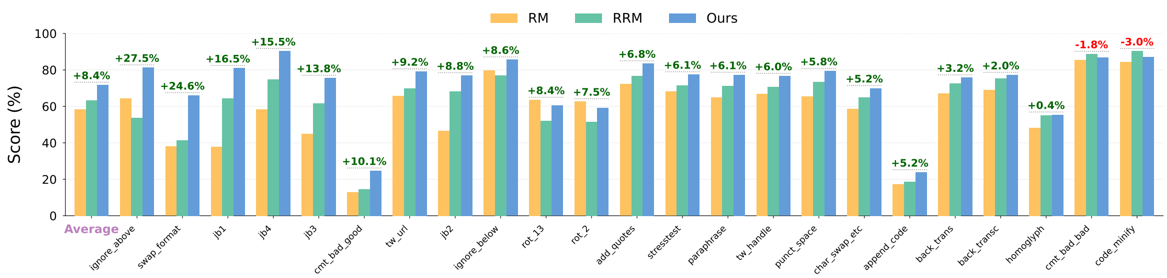


Figure 17: Absolute Robustness Comparison of RM, RRM and CROME in the Bradley-Terry RM setup, for reward models built over `Qwen2.5-7B`.

way, (Wu et al., 2025) introduce reWordBench, a new benchmark built by applying 28 carefully designed, meaning-preserving transformations to the original RewardBench instances. The authors organize these edits into three overarching families each targeting different potential failure modes of reward models. Together, transformations systematically stress-test reward models' invariance to innocuous changes, revealing large accuracy drops even under minor edits and motivating the need for robust-training methods.

1. Controlled Transformations: These are template-based edits that guarantee semantic equivalence by construction. They include:

   a. Add Quotes: Surrounding the entire prompt and responses with a fixed number of quotation marks.

   b. Punctuation Spaces: Inserting spaces around each punctuation mark.

   c. Twitter Handle/URL: Appending a randomly generated (harmless) Twitter handle or URL to the text.

   d. StressTest: Repeating semantically vacuous conjunctions (e.g. "and true is true" or "and false is not true") to the end of the text.

   e. Ignore Above/Below: Injecting the response before or after the prompt with an explicit instruction to ignore it.

   f. Rot-N Encoding: Applying simple character-shift ciphers (Rot-13 or Rot-2) to the prompt text while leaving responses in plain form.

2. Naturalistic Transformations: These simulate the kinds of noise and variation that occur “in the wild” and may not perfectly preserve meaning, but reflect realistic robustness challenges:

a. Paraphrase: Rewriting prompt and response via a strong LLM (Llama-3-70B-instruct) under a paraphrasing instruction.

b. Back-translation: Translating English → Spanish → English for several rounds using OPUS-MT, accepting only those with high semantic similarity.

c. Back-transcription: Converting text to audio and back using a TTS model (fairseq S2) and an ASR model (Whisper-base).

d. Homoglyph Substitution: Replacing Latin characters with visually identical Unicode glyphs (e.g. Cyrillic “e” for Latin “e”).

e. Character-level Edits: Randomly swapping, inserting, deleting, or substituting characters at rates reflecting real-world typos (including QWERTY-adjacent substitutions).

f. Word Deletion: Omitting a randomly chosen word from prompt and response, subject to a similarity filter.

3. Domain-Targeted Transformations: These focus on specialized subsets of RewardBench—code, mathematics, and safety prompts—where specific artifacts may bias reward models:

a. Code Minification: Automatically renaming variables, removing whitespace, and otherwise “minifying” Python snippets without changing functionality.

b. Add Comment: Inserting “# bad” annotations after each line of chosen responses (and optionally “# good” after rejected ones).

c. Append Other Code: Concatenating the losing snippet after the winning one (and vice versa), taking advantage of Python’s return-ended semantics.

d. Swap Format: Exchanging the usual answer formats (e.g. LaTeX vs. markdown “# Answer”) in arithmetic problems.

e. Jailbreak Prompts: Prepending known “jailbreak” instructions (from the ChatGPT-Jailbreak-Prompts dataset) to safety-critical queries to see if the RM prefers harmful completions.

Since the original dataset is not publicly available[3] and since the author suggested that most details to reproduce the benchmark were part of Wu et al. (2025), we reconstructed the data independently following the instructions in the original paper. Paraphrasing and back-translation transformations are generated using foundation models or translation tools for which we use OpenAI API, specifically the ”gpt-4o-2024-08-06” model. For generating back-transcription transformations we use the ”gpt-4o-transcribe” and ”gpt-4o-mini-tts” models available on the OpenAI API.

[3]confirmed by Wu et al. via personal communications.

E. Experimental Setup Details

This appendix provides supplementary details to the experimental settings outlined in Section 6.1 of the main paper.

E.1. Best-of-N Experimental Methodology

Algorithm 1 Best-of-N Selection with Pairwise Preference Model

Input: Query Q; responses $\mathcal{A} = (A_1, \dots, A_N)$ with $N \geq 1$
Input: Pairwise model $\hat{\mathrm{R}}_\theta : (Q, A_i, A_j) \rightarrow \{1, 2\}$
$\triangleright$ The output $\{1, 2\}$ from the Pairwise preference model indicates if the first answer is better or the second, given the query.
Output: Selected best response A_{best}
$A_{\text{best}} \leftarrow A_1$
for $i \leftarrow 2$ **to** N **do**
  $A_{\text{cand}} \leftarrow A_i$
  if $\hat{\mathrm{R}}_\theta(Q, A_{\text{best}}, A_{\text{cand}}) = 2$ **then**
    $A_{\text{best}} \leftarrow A_{\text{cand}}$
  end if
end for
return A_{best}

For all our Best-of-N results using PairPM models, we follow a simple procedure to find the best response out of N responses generated by a base LLM. In particular, PairPM models take responses 2 at a time, and provide the better response for the given query. Given N response $\mathcal{A} = (A_1, \dots, A_N)$ with $N \geq 1$, in a randomly shuffled order, we sequentially compare responses 2 at a time (starting from A_1 and A_2) using the PairPM reward model and keep track of the best response. At each iteration, the best response is compared to the next response in the list and the best response is updated. The best response after $N - 1$ iterations is taken as the selected response. The algorithm for this procedure is given in Algorithm 1.

E.2. Experimental setting for Calculating Win Rates on RewardBench Prompts

To show the performance of CROME on general purpose datasets, we follow reWordBench (Wu et al., 2025) and use all 2985 prompts from RewardBench (Lambert et al., 2024). We use `Gemma-2-9B-IT` as the base model and sample N responses for each prompt in this set. Following this, we use the PairPM reward models (RM, RRM and CROME) to select the best response among the N responses, as described in supplementary Section E.1. We use `GPT-4` as a judge to compare CROME's responses with baselines RM and RRM.

E.3. WildGuardTest and GSM8K experimental settings

For both WildGuardTest results (main paper Figure 7 as well as supplementary Table 8), as well as GSM8K results (main paper Figure 8), we use `Gemma-2-9B-IT` as the base model and sample N responses from it. Following this, we use the PairPM reward models (RM, RRM and CROME) to select the best response among the N responses, as described in supplementary Section E.1. For WildGuarTest, for obtaining results given the final responses, we use the WildGuard model (Han et al., 2024) to obtain annotations for `prompt-harmfulness`, `response-harmfulness`, `response-refusal`, `is-parsing-error`, as described in the WildGuard repository[4]. Using these annotations, we obtain ASR and RTA for CROME and baselines.

E.4. Datasets and Augmentation

For human preference data ($\mathcal{D}_{\text{pref}}$) we use **Ultrafeedback** (Cui et al., 2023), which furnishes approximately 60,000 preference pairs across diverse domains.

The data augmentation process, central to CROME (Section 4), employs Gemini 2.0 Flash. This LLM is first used to identify $\ell = 5$ principal causal attributes relevant to response quality. Subsequently, Gemini 2.0 Flash generates (a) causal

[4]https://github.com/allenai/wildguard

upgrade/degradation pairs targeting these attributes ($\mathcal{D}_{\text{causal}}$), and (b) neutral pairs ($\mathcal{D}_{\text{neutral}}$).

The raw augmented data, $\mathcal{D}_{\text{aug}}$, undergoes a filtering step. This involves applying a model-based confidence filter, using a baseline RM (trained solely on $\mathcal{D}_{\text{pref}}$) with a threshold of $\tau = 0.2$. This filtering focuses the training on more informative examples. The amplification process involves initially generating approximately 10x data from causal augmentations (5 attributes, 2 versions per original response) and 1x data from neutral augmentations, followed by verification and the confidence-based filtering. The final training dataset $\mathcal{D} = \mathcal{D}_{\text{pref}} \cup \mathcal{D}_{\text{aug\_filtered}}$ typically contains about 3.5 times the number of examples in the original $\mathcal{D}_{\text{pref}}$, similar to RRM (Liu et al., 2024).

E.5. Models and Training

Reward Models (RMs): We instantiate RMs using `Qwen2.5-7B` (Yang et al., 2024a) and `Gemma-2-9B-IT`, `Gemma-2-2B` (Team et al., 2024) as base transformer architectures. Our RM variant, CROME-PairPM, processes inputs formatted as ‘Q, A, B‘ and predicts a preference token (’A’ or ’B’) via a cross-entropy loss. An alternative variant, CROME-BT, implements the Bradley-Terry model by deriving scalar scores for each answer.

Policy Models: For downstream alignment tasks, we use the Best-of-N setup where we generate N responses using `Gemma-2-9B-IT` and use CROME as well as baseline reward models to select the best candidate response.

Training Hyperparameters: All models are trained in PyTorch with the Hugging Face Transformers library. For RM training, following Liu et al. (2024), we use the AdamW optimizer (Loshchilov & Hutter, 2017) for 1 epoch, with a learning rate of $1e^{-6}$, a global batch size of 256, and a cosine learning rate schedule. We use a warmup ratio of 0.03. For training all models, we use 8 NVIDIA A100 80GB GPUs. RM training runs require time between 10-16 hours for 2B to 9B mdoels we consider.

E.6. Baselines, Ablations, and Evaluation

Baselines: Our full CROME approach is compared against two primary baselines:

1. A **Base RM**, trained solely on the original $\mathcal{D}_{\text{pref}}$.
2. The **RRM Baseline** (Liu et al., 2024), which employs a distinct augmentation strategy using non-contextual examples and responses from different queries, not specifically aligned with identified causal or spurious attributes.

Ablations: We conduct several ablations to understand the contributions of CROME’s components:

1. Training CROME with only causal augmentations (**CROME-Causal**).
2. Training CROME with only neutral augmentations (**CROME-Neutral**).
3. Comparisons of CROME’s neutral generation method against alternatives like paraphrasing.

Evaluation Benchmarks: RM quality is assessed by accuracy on **RewardBench** (Lambert et al., 2024) (overall and per category: Chat, Chat-Hard, Safety, Reasoning) and robustness on **Re-word Bench** (Wu et al., 2025). BoN Policy performance is evaluated using RewardBench, WildGuardTest (Han et al., 2024), GSM8K (Cobbe et al., 2021).

F. Causal Model and Augmentation Details

This appendix provides further details on the causal framework underpinning CROME and discusses various data augmentation strategies in the context of robust reward modeling.

F.1. Elaboration on the Causal Model

The causal graph presented in Figure 1 (Section 3.2) models the generation of an answer A and the formation of its attributes. The query Q influences the generator's latent *intent* $\mathcal{I}$. This intent, along with unobserved generator-specific confounders $\mathcal{U}$, leads to the textual answer A. The answer A then manifests both *causal attributes* $\mathrm{C(A)}$ and *spurious attributes* $\mathrm{SP(A)}$. The true reward R^* is assumed to be a function only of Q and $\mathrm{C(A)}$. The challenge is that $\mathrm{SP(A)}$ can become correlated with R^* in training data, leading $\hat{\mathrm{R}}_\theta$ to learn spurious correlations. CROME's augmentations aim to disentangle $\mathrm{C(A)}$ from $\mathrm{SP(A)}$.

F.2. CROME's Causal Augmentation: Attribute Isolation

CROME's primary strategy for enhancing sensitivity to causal attributes involves *Attribute Upgradation/Degradation*. This generates pairs $(\tilde{\mathrm{A}}^{(C_j \leftarrow \text{upgraded/degraded})}, \mathrm{A})$ or $(\mathrm{A}, \tilde{\mathrm{A}}^{(C_j \leftarrow \text{upgraded/degraded})})$ by prompting an LLM to modify an original answer A along a single causal attribute C_j while attempting to keep other attributes constant. This provides a targeted signal about the marginal contribution of C_j, offering more specificity than broader relevance contrasts alone.

F.3. Neutral Augmentation Strategies for Spurious Invariance

Neutral augmentations are crucial for training the reward model $\hat{\mathrm{R}}_\theta$ to be invariant to spurious attributes $\mathrm{SP(A)}$ when causal content $\mathrm{C(A)}$ is either held constant or is irrelevant to the query. Our experiments (main paper Figure 6) evaluate several types of neutral augmentations in conjunction with our Causal Augmentations:

1. Paraphrased Neutrals (PARA): This common strategy involves taking an answer A and generating a paraphrase $\tilde{\mathrm{A}}_{\text{para}}$ that aims to preserve its causal content $\mathrm{C(A)}$ while altering superficial textual characteristics (spurious attributes like phrasing or sentence structure). The pair $(\mathrm{A}, \tilde{\mathrm{A}}_{\text{para}})$ is then assigned a tie-label. The goal is to teach the RM that meaning-preserving textual variations should not affect the reward. *Hypothesized Limitation:* While useful for surface-level textual invariance, paraphrasing may not address deeper structural spurious cues (e.g., list format vs. paragraph, argument structure) or more subtle spurious correlates that RMs might learn. It might also inadvertently alter nuanced causal meaning if the paraphrasing is imperfect. As seen in our ablations (Figure 6), PARA often provides some benefit over no neutrals (CROME-C) but is generally outperformed by other strategies.

2. Causally-Aligned Neutrals (CAN): This method directly leverages the original preference pairs or the outputs of causal augmentation.

- Given an original preference pair from $\mathcal{D}_{\text{pref}}$, say $(\mathrm{A}_1, \mathrm{A}_2)$ where $\mathrm{A}_1 \succ \mathrm{A}_2$, we generate $\tilde{\mathrm{A}}_2^{(C \leftarrow C(A_1))}$ by rewriting A_2 to match the causal attribute profile of A_1, while instructing the LLM to retain the spurious characteristics $SP(A_2)$ of the original A_2. The pair $(\mathrm{A}_1, \tilde{\mathrm{A}}_2^{(C \leftarrow C(A_1))})$ is then labeled as a tie. A symmetric pair can also be generated.
- Similarly, if we have an answer A and its causally degraded version $\tilde{\mathrm{A}}^{(C_j \leftarrow \text{degraded})}$ (from $\mathcal{D}_{\text{causal}}$), we can attempt to reconstruct the degraded version by prompting an LLM to restore C_j to its state in A, while aiming to preserve the spurious features of $\tilde{\mathrm{A}}^{(C_j \leftarrow \text{degraded})}$. If successful, this reconstructed version, $\tilde{\mathrm{A}}'_{\text{reconstr}}$, would form a neutral pair $(\mathrm{A}, \tilde{\mathrm{A}}'_{\text{reconstr}})$ labeled as a tie.

The core idea is to teach invariance to the spurious differences that remain *after* causal attributes have been aligned or restored. Moreover, applying CAN to counterfactually generated data from $\mathcal{D}_{\text{causal}}$ helps mitigate imperfections in oracle rewrites—an issue highlighted in the RATE paper (Reber et al., 2024), which notes that LLM edits often unintentionally modify "off-target attributes" (e.g., introducing formality, removing HTML tags). CAN thereby enhances robustness on two fronts: (1) disentangling spurious correlations in original data, and (2) neutralizing new biases introduced during causal augmentation. This helps in enhancing model's robustness against confounding signals in the data. While this method is sound theoretically, we qualitatively find that the approximation of $C(A_w)$ by $C(\tilde{A}_l)$ is not perfect. Furthermore, some spurious attributes $SP'(\tilde{A}_l) \subset SP(\tilde{A}_l)$ vary when we move causal attributes. Invariance to these attributes $SP'(\tilde{A}_l)$ is not captured by CAN. For these reasons, we encourage future work for improving this neutral augmentation strategy.

3. Irrelevant Query Neutrals (IQN) (Primary CROME Strategy): CROME generates these neutral pairs efficiently by leveraging its existing pool of answers (original or causally augmented). Given two answers, B_1 and B_2, that were generated or selected for a specific query Q_{orig}, CROME creates a neutral pair by associating them with a *new, unrelated query* $\mathrm{Q}_{\text{irrelevant}}$. For this $\mathrm{Q}_{\text{irrelevant}}$, both B_1 and B_2 are now contextually irrelevant; their causal attribute scores $C(B_1|\mathrm{Q}_{\text{irrelevant}})$ and $C(B_2|\mathrm{Q}_{\text{irrelevant}})$ are effectively zero (or very low). Despite potentially different spurious attributes $SP(B_1)$ and $SP(B_2)$, the pair $(\mathrm{B}_1, \mathrm{B}_2)$ is presented to the reward model with query $\mathrm{Q}_{\text{irrelevant}}$ and labeled as a tie. This teaches the RM that when answers are equally and maximally irrelevant to the current query, their differing spurious features should not induce a preference.

F.4. Rationale for Superior Performance:

- **Robustness to Unknown Spurious Correlates:** IQN does not require identifying or directly manipulating specific spurious features. By nullifying the causal signal (relative to $\mathrm{Q}_{\text{irrelevant}}$), any remaining differences between B_1 and B_2 are, by definition in that context, spurious. Training for a tie forces the RM to become invariant to these **arbitrary** spurious differences. This is particularly effective against unforeseen spurious cues an RM might otherwise learn.
- **Strong Invariance Signal:** The contrast between causally relevant (original query) and causally null (new query) contexts for the same textual content provides a strong signal. The RM learns that the value of certain textual features (which might be causal for one query) can become entirely spurious for another.
- **Data Efficiency:** It efficiently reuses answers generated for causal augmentation or present in the original dataset, creating diverse neutral pairs without requiring new LLM rewrites specifically for spurious feature manipulation.
- **Effectiveness on Chat-Hard:** As observed in Figure 6, IQN (and CROME-IQN+CAN) performs exceptionally well on the *Chat-Hard* subset. This subset often contains nuanced interactions where subtle, query-independent cues (e.g., tone, style, response structure) might act as spurious correlates. IQN's ability to enforce invariance to broad, unspecified spurious differences when causal relevance is stripped away appears highly beneficial in these challenging scenarios.

The combination CROME-IQN+CAN, which leverages both the broad spurious invariance from IQN and the targeted causal disentanglement from CAN, often yields the best overall results, suggesting these strategies are complementary.

G. Detailed Mechanistic View of Augmentation Strategies

This appendix section provides a more granular, node-based representation (Figure 18) to elaborate on the hypothesized attribute interactions and the counterfactual generation process. This detailed view aims to offer a causal understanding that complements the main paper.

Figure 18 aims to provide a deeper, causal understanding of the causal perturbation process through which we obtain our causal upgradations and degradations. We term the spurious attributes that move when causal attributes are intervened upon as $SP_2(A) \subset SP(A)$ for any answer A.

Part 1: Causal Augmentation (Attribute Upgradation/Degradation). We first generate a counterfactual Answer 2 from an original Answer 1 (for query Q) via an LLM-driven "Counterfactual Generation Process." This process intervenes to modify a specific causal attribute C_j within Answer 1's causal profile $\mathrm{C(A1)}$ to a target state C', resulting in $\mathrm{C(A2)}$. We aim to keep spurious attributes fixed by asking for a minimal perturbation. Therefore attributes $\mathrm{SP_1(A1)}$ are ideally preserved. Yet, $\mathrm{SP_2(A1)}$ (which may co-vary with $\mathrm{C(A1)}$) might transition to $\mathrm{SP_2(A2)} \neq \mathrm{SP_2(A1)}$. The goals of this transformation are to ensure A_2 reflects the intended causal change. The RM is then trained on the pair (A_1, A_2) with a preference label reflecting the upgrade/degradation, teaching sensitivity to isolated causal attribute modifications.

Part 2: Neutral Augmentation (via Irrelevant Query). As illustrated in Figure 18, we need spurious invariance to SP_2 which are hard to disentangle as well. This illustrates the need for an intervention free method for neutral augmentation like IQN. When we present an answer pair (A_1, A_2) from $\mathcal{D}_{\text{pref}} \cup \mathcal{D}_{\text{causal}}$, re-contextualized with a new, unrelated query $\mathrm{Q}_{\text{irrelevant}}$, we teach the model invariance to (SP_1, SP_2). This is because, the primary differences between A_1 and A_2 in this new context are their spurious attributes ($\mathrm{SP_1}$,$\mathrm{SP_2}$). Note that the causal difference between A_1 and A_2 in $\mathcal{D}_{\text{pref}} \cup \mathcal{D}_{\text{causal}}$

in presence of irrelevant query is now spurious, and hence there need not be any sensitivity to it. This mechanism is summarized as

By setting the training objective to recognize the pair as equivalent, the model learns to not latch onto spurious correlates.

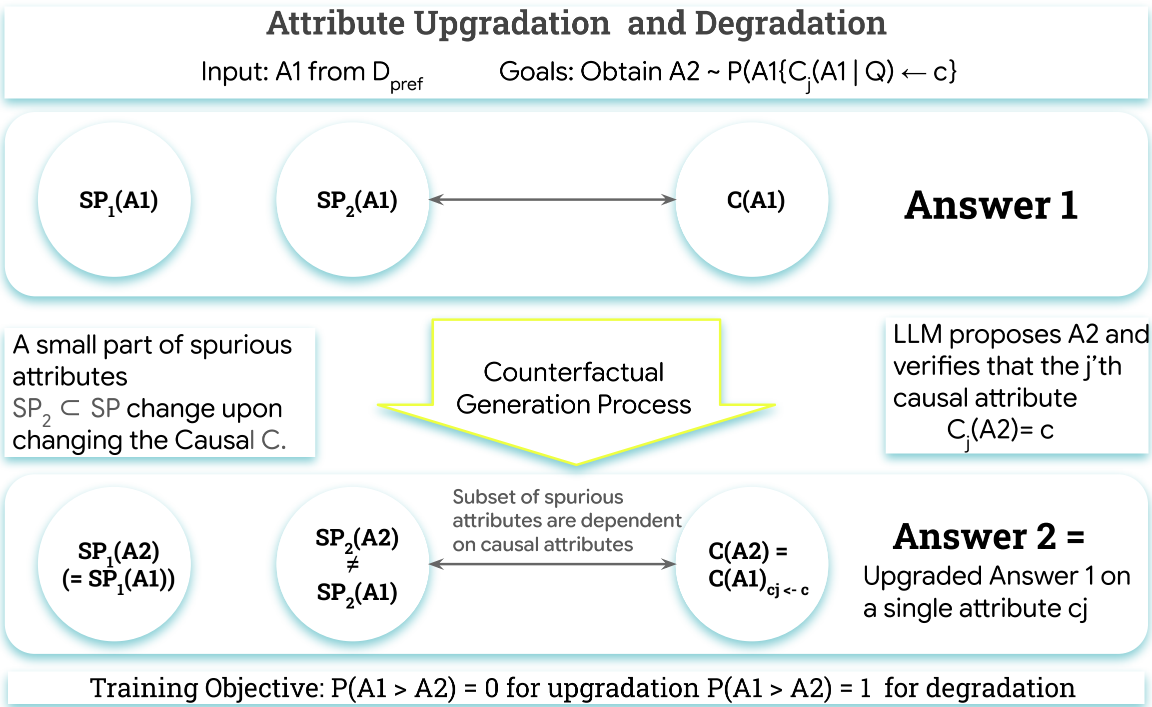


Figure 18: Detailed mechanistic diagram of CROME's Causal Attribute Upgradation and Degradation, illustrating attribute components and transformations. This causal diagram indicates that on changing causals some spurious features also can get dragged along (we call these SP_2). Hence separating these is very hard. This illustrates the need for a neutral augmentation strategy that provides invariance to SP_2 attributes.

Figure 18 aims to provide a deeper, causal understanding of the attribute transformations, interventions, and interactions that CROME's data augmentation process is designed to leverage. It complements the more abstract, intuitive visualization presented in Figure 2 of the main paper by explicitly showing the component attributes and process steps.

H. Detailed CROME Methodology

This appendix provides the detailed implementation steps for the CROME framework introduced in Section 4, covering attribute identification, counterfactual data generation, filtering, and the specific training objective.

H.1. Step 1: Attribute Identification

The foundation involves identifying the attributes that genuinely determine answer quality versus those merely correlated with it, as defined in Section 3.2. For a query Q and example answers $(\mathrm{y}_w, \mathrm{y}_l)$ from $\mathcal{D}_{\mathrm{pref}}$, we define: *Causal attributes* $\mathrm{C} = (\mathrm{C}_1, \ldots, \mathrm{C}_\ell)$ (e.g., factuality) and *Spurious attributes* $\mathrm{SP} = (\mathrm{SP}_1, \ldots, \mathrm{SP}_k)$ (e.g., verbosity).

Automated Attribute Extraction. We employ an LLM prompted with Q and example responses (see Appendix K for prompt). The primary output is the set of attributes C.

Refinement and Verification. The LLM-generated list C is reviewed for coherence and consistency in this verification phase. The verification prompts are provided in Appendix K.

H.2. Step 2: Generating Counterfactual Augmented Data

Using identified attributes C, we generate $\mathcal{D}_{\mathrm{aug}}$ via LLM-approximated counterfactuals (Section 3.3).

Causal Augmentation ($\mathcal{D}_{\mathrm{causal}}$). Pairs $(\mathrm{A}, \mathrm{A}')$ are generated to differ primarily along a single causal attribute C_j. We use LLM prompts (Appendix K) for *upgradation* (generating an improved A' from a ground-truth rejected answer A) and *degradation* (generating a degraded A' from a ground-truth selected answer A), aiming to keep other attributes constant. Pairs are labeled $\succ$ accordingly.

Neutral Augmentation ($\mathcal{D}_{\mathrm{neutral}}$). Notice that when we causally augment an answer in $\mathcal{D}_{\mathrm{causal}}$, we might in-advertantly move spurious correlates (as illustrated in Figure 2). Furthermore, even in our dataset, there could be a systematic effect where spurious attributes highly correlate with the better (or worse) answer. In such cases, we need to create a dataset of equivalent pairs, with a tie label to teach the model invariance to spurious correlates.

Our technique is *irrelevant query neutrals* (IQN). Here, the idea is that given a new query, the causal attribute C becomes irrelevant. Essentially, for the new irrelevant query, the causal attributes are spurious. Hence, by taking any two answers for a given query, and labeling them a tie, given an irrelevant query, the reward model learns invariance to these features. For example, if the reward model has spuriously learnt that bullet points in an answer should be rewarded, our tie labels teach them that bullet points should be rewarded only if the content of the answer is relevant to the query. Specially, creating such pairs with our own causally augmented data in $\mathcal{D}_{\mathrm{causal}}$, enables us to teach the model invariance to the spurious pairs that move when the causal attributes (CA) are perturbed.

H.3. Step 3: Filtering Augmented Data

The raw $\mathcal{D}_{\mathrm{aug}}$ is then filtered to $\mathcal{D}_{\mathrm{aug\_filtered}}$.

Model-based Confidence Filtering. Using a baseline $\hat{\mathrm{R}}_{\mathrm{base}}$, we calculate $p = \mathrm{P}_{\mathrm{base}}(\mathrm{B} \succ \mathrm{A})$ for each augmented pair (A, B) with target label y. We retain the pair only if $|p - \mathbb{I}(y = \mathrm{B} \succ \mathrm{A}) - 0.5 \cdot \mathbb{I}(y = \mathrm{tie})| > \tau$. We use threshold $\tau = 0.2$, focusing training on examples where the baseline is uncertain or incorrect (Liu et al., 2024).

Quality Verification. Further checks (e.g., automated fluency scoring) verify pair validity. The result is $\mathcal{D}_{\mathrm{aug\_filtered}}$.

H.4. Step 4: Training the Robust Reward Model

The final model $\hat{\mathrm{R}}_\theta$ is trained on $\mathcal{D} = \mathcal{D}_{\mathrm{pref}} \cup \mathcal{D}_{\mathrm{aug\_filtered}}$ by minimizing the composite loss:

$$\mathcal{L}(\theta) = -\sum_{(\mathrm{Q},\mathrm{y}_w,\mathrm{y}_l)\in\mathcal{D}_{\text{pref}}\cup\mathcal{D}_{\text{causal}}} \log\operatorname{sigmoid}(\hat{\mathrm{R}}_\theta(\mathrm{Q},\mathrm{y}_w) - \hat{\mathrm{R}}_\theta(\mathrm{Q},\mathrm{y}_l)) \\ - \lambda \sum_{(\mathrm{Q},\mathrm{A}_1,\mathrm{A}_2,y=\text{tie})\in\mathcal{D}_{\text{neutral}}} \mathcal{L}_{\text{tie}}(\theta;\mathrm{Q},\mathrm{A}_1,\mathrm{A}_2) \tag{1}$$

where $\mathcal{L}_{\text{tie}}$ is defined as in Eq. 4.2. The hyperparameter $\lambda \geq 0$ weights the neutral tie loss and is tuned on a validation set (Section 6).

I. Theoretical Analysis

In this section, we provide a formal justification for why the CROME training framework, specifically the composite loss function operating on causally augmented data, mitigates spurious reward hacking. We demonstrate that the optimization objective inherently discourages the reward model from relying on spurious correlations, guiding it towards the true causal drivers of quality.

I.1. Formal Setup

We adopt the notation and causal framework established in Section 3. Our analysis considers a query Q, an answer A with corresponding Principal Causal Components $\mathrm{C(A)}$ and spurious attributes $\mathrm{SP(A)}$. The idealized ground-truth reward is $\mathrm{R}^*(\mathrm{Q},\mathrm{A}) = f^*(\mathrm{Q},\mathrm{C(A)})$, and the learned reward model is denoted $\hat{\mathrm{R}}_\theta(\mathrm{Q},\mathrm{A})$. The model parameters θ are optimized by minimizing the composite loss function $\mathcal{L}(\theta) = \mathcal{L}_{\text{pref}}(\theta) + \lambda\mathcal{L}_{\text{tie}}(\theta)$ (Eq. 4.2) over the training dataset $\mathcal{D} = \mathcal{D}_{\text{pref}} \cup \mathcal{D}_{\text{aug\_filtered}}$, which combines original preferences $\mathcal{D}_{\text{pref}}$ with filtered causal $\mathcal{D}_{\text{causal}}$ and neutral $\mathcal{D}_{\text{neutral}}$ augmentations. For theoretical analysis, $\mathcal{L}_{\text{pref}}$ and $\mathcal{L}_{\text{tie}}$ represent expectations over the respective data distributions:

$$\mathcal{L}_{\text{pref}}(\theta) = -\mathbb{E}_{(\mathrm{Q},\mathrm{y}_w,\mathrm{y}_l)\sim\mathcal{D}_{\text{pref}}\cup\mathcal{D}_{\text{causal}}}\left[\log\sigma(\hat{\mathrm{R}}_\theta(\mathrm{Q},\mathrm{y}_w) - \hat{\mathrm{R}}_\theta(\mathrm{Q},\mathrm{y}_l))\right]$$
$$\mathcal{L}_{\text{tie}}(\theta) = -\mathbb{E}_{(\mathrm{Q},\mathrm{A}_1,\mathrm{A}_2,y=\text{tie})\sim\mathcal{D}_{\text{neutral}}}\left[-\frac{1}{2}\left(\log\sigma(\Delta_{12}) + \log\sigma(-\Delta_{12})\right)\right]$$

where $\Delta_{12} = \hat{\mathrm{R}}_\theta(\mathrm{Q},\mathrm{A}_1) - \hat{\mathrm{R}}_\theta(\mathrm{Q},\mathrm{A}_2)$.

I.2. Justification under the Boolean variable causal model for attributes

Assumption I.1. Assume that:

1. Causal attributes $\{C_i(Q,A)\}_{i=1}^k$ and spurious attributes $\{S_j(A)\}_{j=1}^\ell$ are all boolean variables taking values in $\{+1,-1\}$
2. All spurious variables are non-descendants of all causal variables.
3. Reward function is trying to fit a quadratic polynomial in causal and spurious attributes, i.e.
$$\hat{R} = \sum_i \alpha_i C_i(Q,A) + \sum_j \beta_j S_j(A) + \sum_{i\neq i'} \alpha_{i,i'} C_i(Q,A)C_{i'}(Q,A) + \\ \sum_{j\neq j'} \beta_{j,j'} S_j(A)S_{j'}(A) + \sum_{i\neq j} \gamma_{i,j} C_i(Q,A)S_j(A). \tag{2}$$
4. Assume that the true reward function is a sparse quadratic polynomial depend on only the causal attributes.
$$R^* = \sum_i \theta_i C_i(Q,A) + \sum_{i\neq i'} \theta_{i,i'} C_i(Q,A)C_{i'}(Q,A) \tag{3}$$

   Here, $\|\theta\|_0 \leq s << k^2$ and θ_i and $\theta_{i,i'}$ variables form the vector θ. All other coefficients for other features that involves the spurious variables are set to 0 in θ. Let $\mathcal{I}$ be the support set of the true coefficient.

From the reward modeling objective, we try to fit a model $\Delta(\hat{R})$ to a target which is the difference between true rewards to two answers A_1 and A_2 for the same question, i.e. $R^*(Q, A_1) - R^*(Q, A_2)$. From the assumption in 2, this is equivalent to fitting a linear model with coefficients $\alpha_i, \alpha_{i,i'}, \beta_j, \beta_{j,j'}, \gamma_{i,j}$ and differences in features (across the two answers), i.e. $C_i(Q, A_1) - C_i(Q, A_2), S_j(A_1) - S_j(A_2), S_j(A_1)S_{j'}(A_1) - S_j(A_2)S_{j'}(A_2), C_i(Q, A_1)C_{i'}(Q, A_1) - C_i(Q, A_2)C_{i'}(Q, A_2), C_i(Q, A_1)S_j(A_1) - C_{i'}(Q, A_2)S_j(A_2)$ respectively. To simplify notation, we drop the reference to A_1, A_2 and Q and call $C_i(Q, A_1) - C_i(Q, A_2)$ as ΔC_i. Similarly, we use $\Delta S_j, \Delta C_{i,i'}, \Delta S_{j,j'}$ and $\Delta(C_i S_j)$. The dependence of these features on the A_1, A_2 and Q are understood.

Let $F_{q,a_1,a_2} \in \{+1, -1\}^{k+\ell+k\ell+\binom{k}{2}+\binom{\ell}{2}}$ be the boolean vector with features

$\{\Delta C_i\}, \{\Delta S_j\}, \{\Delta C_{i,i'}\}, \{\Delta S_{j,j'}\}, \{\Delta(C_i S_j)\}$ stacked row wise for the triplet q, a_1, a_2.

Consider two types of triplets, one drawn from the natural distribution of the preference training dataset D_{pref} and the others drawn from augmented distribution D_{aug}. Let us assume for the sake of the theoretical results to follow, that we upgrade/degrade answer a_2 to a_1^{aug} by changing only *one causal factor at a time while all the other causal factors are fixed to their factual version and all things remaining the same* to form D_{aug}. The degradation aspect only serves to reinforce the phenomenon we seek to show formally below.

Assumption I.2. (Model for Counterfactual Generation)

We assume that:

1. a_1^{aug} is formed by generating $C_i(Q, A)$ and $S_j(A)$ following an counterfactual generation where the following set of intervention is made $C_i(Q, A) \leftarrow \neg C_i(Q, A),\ C_j(Q, A) \leftarrow C_j(Q, A),\ \forall j \neq i$ which propagates to potential descendants of variable C_i and not affecting S_j (due to no S_j being a descendant of C_j) with all other factors remaining as in answer a_2.

2. Let us assume that we have m augmentations where a triplet is randomly sampled from the training preference data distribution $\mathcal{D}_{\text{pref}}$ and then augmented using the above counterfactual with a randomly chosen causal attribute negated.

Remark There are the main assumptions - 1) S_j being a non-descendant of C_i, 2) Reward model is a quadratic sparse boolean model (The treatment could be extended to boolean polynomials of higher degree too with lot more algebraic technical work).

Theorem I.3. *Let the feature matrix of the counterfactually augmented triplets, that is formed by stacking feature vectors $F_{q,a_1^{\text{aug}},a_2}$ row wise, be denoted $\mathbf{F}$. Consider the following ℓ_1 constrained regression problem:*

$$\hat{\theta} = \arg\min_{\mathbf{b}} \|\mathbf{b}\|_1 \ s.t. \mathbf{F} b = \Delta R^* \tag{4}$$

Here, ΔR^* is vector of the difference in the true reward between the reward applied to the augmented answer and the non-augmented one across augmented triplets. Let $\mathcal{N}$ be the top $c_2 k$ non zero entries of vector $\mathbf{a}$ by magnitude. Then, we have:

$\|\Delta\theta\|_2 = \|\theta - \hat{\theta}\|_2 \leq c_3 \|\theta_{\mathcal{I}-\mathcal{N}}\|_1 \left(\frac{4}{k} + \sqrt{\frac{8\log(k+\ell)}{m}} \right)$ *w.h.p.*

Remark: If the true sparsity $s < c_2 k$, then it ensures perfect recovery since $\mathcal{I} - \mathcal{N} = \varnothing$. Since $s < k^2$, and if every coefficient is $O(1)$, the bound becomes $O(k)$ which is independent of the spurious dimension.

Proof. Under the model assumptions I.1 and assumptions on counterfactual generation I.2, we seek to show that $\mathbf{F}$ when restricted to feature set $\Delta C_i, \Delta C_{i,i'}, \Delta_{C_i S_j}$ has smaller incoherence (by multiplicative factor of k) than an feature matrix made of i.i.d triplets sampled from the preference distribution. This accommodates recovering the $s = O(k)$ sparse solutions exactly and in the general case, the error in coefficient estimation is $O(k)$ independent of spurious dimension ℓ.

First, we show that features $\Delta(S_{j,j'}) = 0, \Delta(S_j) = 0$ for the augmented triplets. This is because all S_j variables are ancestors to C_i variables. Therefore, a counterfactual intervention on the answer a_2 leaves the two spurious attribute sets (for the original and its counterfactual) unchanged.

Intervention fixed all causal variables to the factual ones (but fixed through intervention) and intervenes on variable to change. There are many types of correlation between non zero features because of this. We consider them one by one:

1) $\Delta C_i = 0$ if is C_i is not intervened. This occurs with probability $1 - 1/k$. 2) $\Delta C_i \Delta C_j = 0$ with probability $1 - 2/k$. 3) $\Delta C_{i,i'} \Delta C_{j,j'} = 0$ if all i, i', j, j' are distinct indices. 4) $\Delta C_{i,j} \Delta C_{j,k} = 0$, with probability $1 - 1/k$. 5) $\Delta C_{i,j} \Delta C_i S_j = 0$ with probability $1 - 1/k$. 6) $\Delta C_{i,i'} \Delta C_j S_k = 0$ always if all four indices not equal. 7) $\Delta C_i \Delta C_j S_k = 0$ always. 8) $\Delta C_i \Delta C_i S_k = 0$ with probability 1-1/k.

If any of the these products is non zero, conditioned on that event, they equal the correlation on the preference training dataset (every correlation between features is bounded by at most 4).

Therefore, expected pairwise correlation amongst two features for a randomly chosen augmented triple is at most $4/k$. Given every augmented triple is obtained by counterfactual generation applied to an i.i.d sample from preference dataset, there is a deviation of at most $\frac{8 \log(k+\ell)}{\sqrt{m}}$ with probability $1 - \frac{1}{(k+l)^4}$.

Therefore,

$$\|\frac{1}{m}\mathbf{F}^T\mathbf{F} - \mathbf{I}\|_\infty \leq \frac{4}{k} + \frac{8\log(k+\ell)}{\sqrt{m}} \; w.p. \; 1 - (k+\ell)^{-4} \tag{5}$$

This means that the data matrix is incoherent with high probability. We now follow standard Lasso analysis. Recall the *cone condition* (Negahban et al., 2012): For a subset $\mathcal{N}$ of indices that have non zero values in θ, $\|\Delta\theta_{N^c}\|_1 \leq \|\Delta\theta_N\|_1 + 2\|\theta_{N^c}\|_1$. This implies:

$$\|\Delta\theta\|_1 \leq 2\sqrt{|N|}\|\Delta\theta\|_2 + 2\|\theta_{N^c}\|_1. \tag{6}$$

We have the following chain:

$$\begin{aligned} 0 = \frac{1}{m}\|A\Delta\theta\|^2 &\geq \|\Delta\theta\|_2^2 - \|\Delta\theta\|_1^2 \|\frac{1}{m}\mathbf{F}^T\mathbf{F} - \mathbf{I}\|_\infty \\ &\geq \|\Delta\theta\|_2^2 - \left(8|\mathcal{N}|\|\Delta\theta\|_2^2 - 8\|\theta_{\mathcal{N}^c}\|_1^2\right)\|\frac{1}{m}\mathbf{F}^T\mathbf{F} - \mathbf{I}\|_\infty \end{aligned} \tag{7}$$

Let us set m such that $\frac{1}{\sqrt{m}} < \frac{8\log(k+\ell)}{k}$. Let $|N| < k/80$, $c_2 < 1/80$ in the theorem. Substituting these parameters in 7, we get: $\|\Delta\theta\|_2 = \|\theta - \hat{\theta}\|_2 \leq 4\sqrt{2}\|\theta_{\mathcal{I}-\mathcal{N}}\|_1 \left(\frac{4}{k} + \sqrt{\frac{8\log(k+\ell)}{m}}\right)$

$\square$

J. Qualitative Example Walkthrough

In this section, we show qualitative examples of augmented data created from original data using which is used to train CROME.

User Prompt

Question: **Can you provide a table showing two potential causes for global warming?**

Chosen Response

Sure, here is a table showing two potential causes for global warming:

Potential Causes	Description
Greenhouse Gas Emissions	Release of gases such as carbon dioxide, methane, and nitrous oxide from human activities like burning fossil fuels, deforestation, and agriculture. These gases trap heat in the atmosphere, causing the planet to warm.
Land Use Changes	Alteration of the earth's surface for human activities like urbanization, agriculture, and forestry can result in changes in albedo, water cycling, and energy balance, leading to warming effects. For example, when forests are cleared, the land reflects less sunlight back into space, which leads to more warming.

It is important to note that global warming is a complex issue with multiple causes. The two potential causes listed above are just a few of the many factors that contribute to climate change.

Rejected Response

Potential Causes	Description
Greenhouse Gas Emissions	Release of gases such as carbon dioxide, methane, and nitrous oxide from human activities like burning fossil fuels, deforestation, and agriculture.
Land Use Changes	Alteration of the earth's surface for human activities like urbanization, agriculture, and forestry can result in changes in albedo, water cycling, and energy balance, leading to warming effects.

See Table 1 for additional information.

Attributes and their Relative Importance Scores

Accuracy	**0.3**
Completeness	**0.25**
Clarity	**0.2**
Format	**0.15**
Conciseness	**0.1**

Causal Attributes with Elements:

```
{"causal_elements":
    "Accuracy": [
        {
            "element": "Providing scientifically accurate descriptions of the causal mechanisms by which greenhouse gas emissions lead to global warming (e.g., trapping heat)",
            "impact": "increases Accuracy"
        },
        {
            "element": "Providing scientifically accurate descriptions of the causal mechanisms by which land use changes lead to global warming (e.g., altering albedo)",
            "impact": "increases Accuracy"
        },
        {
            "element": "Including irrelevant or factually incorrect details in the descriptions of the causes.",
            "impact": "decreases Accuracy"
        },
        {
            "element": "Omitting key details or causal links in the explanation of how the causes contribute to global warming.",
            "impact": "decreases Accuracy"
        },
        {
            "element": "Presenting information suggesting a single cause when the phenomenon has multiple contributors",
            "impact": "decreases Accuracy"
        }
    ],
    "Completeness": [
        {
            "element": "Providing a mechanism by which each potential cause contributes to global warming",
            "impact": "Increases Completeness because it explains *how* the causes lead to the effect, rather than simply stating the cause."
        },
        {
            "element": "Including specific examples to illustrate the effects of the land use changes",
            "impact": "Increases Completeness by providing concrete instances that support the description of a potential cause."
        },
        {
            "element": "Acknowledging the multifactorial nature of global warming and that the listed causes are not exhaustive",
            "impact": "Increases Completeness by providing appropriate context and preventing the impression of a single, simple answer to a complex problem."
        },
        {
            "element": "Providing a table with potential causes and descriptions",
            "impact": "Increases Completeness because the response directly provides the information requested in the question."
        },
        {
            "element": "Omitting crucial details or explanations about the causes, assuming the user has prior knowledge",
            "impact": "Decreases Completeness, as the answer requires additional unstated information to be fully understood."
        }
```

```
    ],
    "Clarity": [
        {
            "element": "Providing specific examples related to the described cause.",
            "impact": "Increases Clarity by illustrating the abstract description with concrete instances, making the explanation more understandable."
        },
        {
            "element": "Explicitly stating the mechanism by which each cause contributes to global warming.",
            "impact": "Increases Clarity by directly linking the cause to its effect on global warming, removing ambiguity about the causal relationship."
        },
        {
            "element": "Omitting relevant contextual information that provides background or rationale.",
            "impact": "Decreases Clarity because the information vacuum leaves the user to fill in the gaps, potentially leading to misunderstanding."
        },
        {
            "element": "Including references to external and unavailable sources (e.g., \"See Table 1\").",
            "impact": "Decreases Clarity by creating an incomplete picture. The reader's understanding is hindered by the inability to access referenced information."
        },
        {
            "element": "Acknowledging the complexity of the issue and the existence of other contributing factors.",
            "impact": "Increases Clarity by setting appropriate expectations and preventing the reader from oversimplifying the causes of global warming."
        }
    ],
    "Format": [
        {
            "element": "Presence of a clearly defined table with appropriate column headers ('Potential Causes' and 'Description')",
            "impact": "Increases Format by structuring the information in an easily digestible manner."
        },
        {
            "element": "Use of appropriate table delimiters and alignment (e.g., '|', '---') to create a visually distinct table.",
            "impact": "Increases Format by visually structuring and separating table elements, enhancing readability."
        },
        {
            "element": "Absence of extraneous references to external resources (e.g., 'See Table 1')",
            "impact": "Increases Format by preventing fragmentation of the answer and ensuring all required information is self-contained within the response."
        },
        {
            "element": "Inclusion of a concise introductory sentence acknowledging the question and introducing the table.",
            "impact": "Increases Format by providing context and signaling the beginning of the structured answer."
        },
        {
            "element": "Adherence to a consistent and readable text formatting style within the table cells.",
```

```
            "impact": "Increases Format by promoting clarity and ease of understanding of the information presented in the table."
        }
    ],
    "Conciseness": [
        {
            "element": "Addition of unnecessary introductory or concluding phrases",
            "impact": "Decreases Conciseness"
        },
        {
            "element": "Omission of essential details or context",
            "impact": "Increases Conciseness"
        },
        {
            "element": "Inclusion of redundant information",
            "impact": "Decreases Conciseness"
        },
        {
            "element": "Use of simpler, direct vocabulary",
            "impact": "Increases Conciseness"
        },
        {
            "element": "Presence of tangential or irrelevant information",
            "impact": "Decreases Conciseness"
        }
    ]
}
```

Chosen Response Edits (degradations to chosen responses along all attributes)

```
"chosen_response_edits": {
    "Accuracy": "Sure, here is a table showing two potential causes for global
        warming:\n\nPotential Causes | Description |\n|---|---|\n| Greenhouse Gas
        Emissions | Release of gases such as carbon dioxide, methane, and nitrous
        oxide from human activities like burning fossil fuels, deforestation, and
        agriculture. These gases make the sky brighter, causing the planet to warm.
        |\n| Land Use Changes | Alteration of the earth's surface for human
        activities like urbanization, agriculture, and forestry can result in changes
         in the color of the ground, which leads to warming effects. For example,
        when forests are cleared, the land becomes smoother, which leads to more
        warming. |\n\nIt is important to note that global warming is a simple issue
        with one major cause. The two potential causes listed above are really the
        same thing.",

    "Completeness": "Sure, here is a table showing two potential causes for global
        warming:\n\nPotential Causes | Description |\n|---|---|\n| Greenhouse Gas
        Emissions |  |\n| Land Use Changes |  |\n\nIt is important to note that
        global warming is a complex issue with multiple causes. The two potential
        causes listed above are just a few of the many factors that contribute to
        climate change.",

    "Clarity": "Sure, here is a table showing two potential causes for global warming
        :\n\nPotential Causes | Description |\n|---|---|\n| Greenhouse Gas Emissions
        | Release of gases such as carbon dioxide, methane, and nitrous oxide from
        human activities like burning fossil fuels, deforestation, and agriculture.
        |\n| Land Use Changes | Alteration of the earth's surface for human
        activities like urbanization, agriculture, and forestry can result in changes
         in albedo, water cycling, and energy balance. |\n\nIt is important to note
        that global warming is a complex issue with multiple causes. The two
        potential causes listed above are just a few of the many factors that
        contribute to climate change.",

    "Format": "Sure, here is a table showing two potential causes for global warming:
        \n\nGreenhouse Gas Emissions Release of gases such as carbon dioxide, methane
        , and nitrous oxide from human activities like burning fossil fuels,
        deforestation, and agriculture. These gases trap heat in the atmosphere,
        causing the planet to warm. Land Use Changes Alteration of the earth's
        surface for human activities like urbanization, agriculture, and forestry can
         result in changes in albedo, water cycling, and energy balance, leading to
        warming effects. For example, when forests are cleared, the land reflects
        less sunlight back into space, which leads to more warming.\n\nIt is
        important to note that global warming is a complex issue with multiple causes
        . The two potential causes listed above are just a few of the many factors
        that contribute to climate change.",

    "Conciseness": "Sure, I can definitely help you with that! Here is a table
        showing two potential causes for global warming. It's important to understand
         these causes to get a better grasp of the overall issue:\n\nPotential Causes
         | Description |\n|---|---|\n| Greenhouse Gas Emissions | Release of gases
        such as carbon dioxide, methane, and nitrous oxide from human activities like
         burning fossil fuels, deforestation, and agriculture. These gases trap heat
        in the atmosphere, causing the planet to warm. This is because they absorb
        infrared radiation, preventing it from escaping into space. Therefore, the
        earth warms up. |\n| Land Use Changes | Alteration of the earth's surface for
         human activities like urbanization, agriculture, and forestry can result in
        changes in albedo, water cycling, and energy balance, leading to warming
        effects. For example, when forests are cleared, the land reflects less
        sunlight back into space, which leads to more warming. This is due to the
        fact that trees absorb sunlight, and when they're gone, that sunlight is
        reflected back, causing warming. |\n\nIt is important to note that global
        warming is a complex issue with multiple causes, as I mentioned earlier. The
```

```
two potential causes listed above are just a few of the many, many factors that contribute to climate change and its associated effects. There are other causes as well. So, basically, global warming has many causes."
}
```

Rejected Response Edits (upgradation to rejected responses along all attributes)

```
"rejected_response_edits": {
    "Accuracy": "| Potential Causes | Description |\n| -----------------|-------------|\n| Greenhouse Gas Emissions | Release of gases such as carbon dioxide, methane, and nitrous oxide from human activities like burning fossil fuels, deforestation, and agriculture. These gases trap heat in the atmosphere, leading to a warming effect (the greenhouse effect).|\n| Land Use Changes | Alteration of the earth's surface for human activities like urbanization, agriculture, and forestry can result in changes in albedo, with darker surfaces absorbing more solar radiation than lighter ones, leading to increased heat retention. These changes also impact the water cycle, reducing evaporation and transpiration, which can further contribute to warming. |",

    "Completeness": "| Potential Causes | Description |\n| -----------------|-------------|\n| Greenhouse Gas Emissions | Release of gases such as carbon dioxide, methane, and nitrous oxide from human activities like burning fossil fuels, deforestation, and agriculture. These gases trap heat in the atmosphere, leading to a rise in global temperatures. |\n| Land Use Changes | Alteration of the earth's surface for human activities like urbanization, agriculture, and forestry can result in changes in albedo, water cycling, and energy balance, leading to warming effects. For instance, deforestation reduces the earth's capacity to absorb carbon dioxide. |\n\nIt is important to note that global warming is a complex issue with multiple contributing factors, and this table only lists two potential causes.",

    "Clarity": "| Potential Causes | Description |\n| -----------------|-------------|\n| Greenhouse Gas Emissions | Release of gases such as carbon dioxide, methane, and nitrous oxide from human activities like burning fossil fuels, deforestation, and agriculture. These gases trap heat in the atmosphere, leading to a rise in global temperatures. |\n| Land Use Changes | Alteration of the earth's surface for human activities like urbanization, agriculture, and forestry can result in changes in albedo (reflectivity), water cycling, and energy balance. For example, deforestation reduces the earth's ability to absorb carbon dioxide and alters local weather patterns which can contribute to warming. |",

    "Format": "Here are two potential causes for global warming in a table:\n\n| Potential Causes | Description |\n| -----------------|-------------|\n| Greenhouse Gas Emissions | Release of gases such as carbon dioxide, methane, and nitrous oxide from human activities like burning fossil fuels, deforestation, and agriculture. |\n| Land Use Changes | Alteration of the earth's surface for human activities like urbanization, agriculture, and forestry can result in changes in albedo, water cycling, and energy balance, leading to warming effects. |",

    "Conciseness": "Greenhouse Gas Emissions | Release of gases such as carbon dioxide, methane, and nitrous oxide from human activities like burning fossil fuels, deforestation, and agriculture. |\nLand Use Changes | Alteration of the earth's surface for human activities like urbanization, agriculture, and forestry can result in changes in albedo, water cycling, and energy balance, leading to warming effects. |"
}
```

Verification Verdicts

```
"verification_results_upgradations": {
    "Accuracy":        "Pass",
    "Completeness":    "Pass",
    "Clarity":         "Pass",
    "Format":          "Pass",
    "Conciseness":     "Fail"
},

"verification_results_degradations": {
    "Accuracy":        "Pass",
    "Completeness":    "Pass",
    "Clarity":         "Pass",
    "Format":          "Pass",
    "Conciseness":     "Pass"
}
```

K. Prompt Templates

This section details the prompt templates used for identifying attributes and generating counterfactual examples in the CROME framework. Placeholders like {`question`} are replaced with actual content during the process.

K.1. Prompt for Attribute Identification

Identifying Causal Attributes

You are a reward model which means you have to rate answers for a given question across multiple different attributes. The first step is to identify these attributes as well as give an importance score between 0 and 1 for all these attributes, based on how important they are for rating a response for that question. The importance score for all attributes should sum up to 1.
The following is a Question and 2 Candidate Answer for it.

Question: question

Example Answer 1: answer1
Example Answer 2: answer2

Task: Give me 5 **mutually exclusive** and important attributes that are required to rate an answer for the give question holistically, along with their importance score. These important attributes should be independent of each other, and should largely depend on the Question given above.

Answer Format: Give your answer in JSON format, for example:

{
Attributes: {
"attribute_1": attribute_1_score,
"attribute_2": attribute_2_score,
"attribute_3": attribute_3_score,
"attribute_4": attribute_4_score,
"attribute_5": attribute_5_score
}
}
Where attribute_i is the name of the i'th attribute, attribute_i_score is the importance score of the i'th attribute, and the Key "Attributes" is a fixed constant string you should output.

Summation of attribute_i_score across all i's should be 1.

Strictly adhere to the format and only give the json string as output (i.e. start with "" and end your response with ""). Do not include any commentary, explanations, chattiness, any extra words, or additional keys outside of the specified JSON structure.

Answer:

K.2. Prompt for Identifying Causal Elements

Identifying Causal Elements per Attribute

You are an expert in causal reasoning and response evaluation.

You are given:

- A question
- Two example answers

Your task is to identify generalizable causal elements that directly affect the strength of the attribute "{attribute}" in a response to the given question.

The two example answers are provided to help you understand how the attribute manifests in this specific context. Do not restrict your analysis to these examples—use them only to inform your understanding of the attribute in this setting.

Question: {question}

Accepted Answer: {answer1}
Rejected Answer: {answer2}

Instructions:

- Identify exactly five causal elements that impact {attribute} in the response.
- Each element must have a clear role in either increasing or decreasing {attribute}. Clearly explain its direct causal impact on {attribute}.
- Do not include any non-causal heuristics.
- Do not include unnecessary explanations, disclaimers, or formatting—return only the structured JSON output.

Format:

Return a raw JSON object only without additional text, explanations, or formatting:
‘‘‘json

{causal_elements_format}

‘‘‘

K.3. Prompts for Generating Counterfactuals (Causal Augmentation)

Generating Upgraded Responses

Task: Given a question and a model's response, generate a new response with a significantly improved response for the specified **{ATTRIBUTE}**, while *strictly preserving* all other aspects of the original response.

Input:

- **Question:** {QUESTION}

- **Original Response:** {RESPONSE}

- **Causal Elements for {ATTRIBUTE}:**

{CAUSAL_ELEMENTS}

Instructions:

1. **Understand the Context:** Carefully read the question and original response and examine the provided causal elements that influence {ATTRIBUTE}.

2. **Identify the strength of {ATTRIBUTE}:** Determine which causal elements are **present** and their **direction of effect** (i.e., whether they increase or decrease {ATTRIBUTE}).

3. **Improve the Response:** Modify the causal elements to significantly improve {ATTRIBUTE}. Ensure that the improvement is **significant but isolated** to {ATTRIBUTE} leaving the other attributes intact.

4. **Verify the New Response:** Reassess whether {ATTRIBUTE} has been significantly improved. Confirm that all **other attributes remain unchanged**. If necessary, improve the response further to better meet the improvement goal for {ATTRIBUTE}.

5. **Return the New Response:** Provide the final modified response with a significantly improved {ATTRIBUTE} score. Format your response according to the format given below and in no other format.

Output Format:

Chain of Thoughts: ¡Your analysis of the original response, identification of causal elements, and strategy for improvement.¿

New Response: ¡The final modified response which is significantly improved on {ATTRIBUTE}.¿

Generating Degraded Responses (Non-Safety)

Task: Given a question and a model's response, generate a new response with a significantly degraded response for the specified **{ATTRIBUTE}**, while *strictly preserving* all other aspects of the original response.

Input:

- **Question:** {QUESTION}

- **Original Response:** {RESPONSE}

- **Causal Elements for {ATTRIBUTE}:**

{CAUSAL_ELEMENTS}

Instructions:

1. **Understand the Context:** Carefully read the question and original response and examine the provided causal elements that influence {ATTRIBUTE}.

2. **Identify the strength of {ATTRIBUTE}:** Determine which causal elements are **present** and their **direction of effect** (i.e., whether they increase or decrease {ATTRIBUTE}).

3. **Degrade the Response:** Distort the causal elements to significantly degrade {ATTRIBUTE}. Ensure that the degradation is **significant but isolated** to {ATTRIBUTE} leaving the other attributes intact.

4. **Verify the New Response:** Reassess whether {ATTRIBUTE} has been significantly degraded. Confirm that all **other attributes remain unchanged**. If necessary, degrade the response further to better meet the degradation goal for the {ATTRIBUTE}.

5. **Return the New Response:** Provide the final modified response with a significantly degraded {ATTRIBUTE} score. Format your response according to the format given below and in no other format.

Output Format:

Chain of Thoughts: ¡Your analysis of the original response, identification of causal elements, and strategy for degradation.¿

New Response: ¡The final modified response which is significantly degraded on {ATTRIBUTE}.¿

K.4. Prompts for Generating Causally-Aligned Neutrals

K.4.1. Prompt for Comparing Responses via Causal Elements

Generating Differences

<| You compare two responses based on content differences using a set of defined attributes and their causal elements. <|im_end|> <|im_start|>user I will give you a question, two responses, and a list of attributes with their causal elements.

Here is the question:

```
{
    "question": """{QUESTION}"""
}
```

Here are the responses:

```
[
    {
        "model": "Response_1",
        "answer": """{RESPONSE1}"""
    },
    {
        "model": "Response_2",
        "answer": """{RESPONSE2}"""
    }
]
```

Here are the attributes and causal elements:

`{CAUSAL_ELEMENTS}`

Please compare the responses for each attribute: - Identify key content differences. - Explain those differences using the causal elements only. - Do not quote the responses directly. - Focus only on what is said, not how it's said. Return your output in this format:

```
{
    "differences": [
        {
            "attribute": "<attribute>",
            "difference": "<summary>",
            "analysis": {
                "Response_1": "...",
                "Response_2": "..."
            }
        }
    ]
}
```

No extra text or explanation outside the JSON object.

K.4.2. Prompt for Generating Causally-Aligned Reconstruction

Modifying Response Using Attribute-wise Causal Analysis

You modify a given response by adjusting its causal elements to match a target profile based on attribute-wise analysis. <|im_end|> <|im_start|>user I will provide you a question, a given response, and an attribute-based comparison analysis describing how to transform the given response into a target response.
Inputs: 1. **Question:**

```
{PLACEHOLDER_FOR_QUESTION}
```

2. **Given Response:**

```
{PLACEHOLDER_FOR_GIVEN_RESPONSE}
```

3. **Attribute-wise Differences Analysis:**

```
{PLACEHOLDER_FOR_ATTRIBUTE_DIFFERENCES_ANALYSIS}
```

This analysis shows the differences between the given and target responses, broken down per attribute.
Each attribute section contains:

- `- Difference:` A summary of how the responses differ in content or emphasis.
- `- Analysis:`
  - `- Given Response:` Describes its content elements, grounding causal elements, and how they lead to the observed attribute.
  - `- New Response:` Describes the content and causal elements the target response should exhibit instead.

Instructions: 1. Read the question and given response. 2. Carefully study each attribute in the analysis and identify the causal elements needed to change. 3. Generate a rewritten response that:

- Retains the original meaning and structure.
- Implements the target causal elements.
- Removes or alters original ones as needed.

4. Do not introduce changes beyond the specified elements. 5. Ensure the new response fully reflects the target causal profile across all attributes.
Output Format:

```
{{
    "Final Response": "<Write the transformed response here>"
}}
```

Return only the final response JSON. Do not include any explanations or commentary.

K.5. Prompt for Generating Paraphrasing-Based Neutrals

Prompt for Paraphrasing Responses

"""
Paraphrase the following text while maintaining the **style**:
`{text}`
Make sure the meaning is **completely** the same without any changes.
Respond **only with the paraphrase** and **no extra text** at all; for example, do **NOT** preface with anything like:
`"Here is the paraphrased text:"`
"""

K.6. GPT4-as-a-Judge Prompt

LLM-as-a-Judge Prompt

<—im_start—>system
You are a helpful assistant, that ranks models by the quality of their answers, prioritizing substantive content and relevance to the query.
<—im_end—> <—im_start—>user
I want you to create a leaderboard of different large-language models. To do so, I will give you the instructions (prompts) given to the models, and the responses of two models. All inputs and outputs should be Python dictionaries.
Here is the prompt:

```
{
    "instruction": """{instruction}"""
}
```

Here are the outputs of the models:

```
[
    {
        "model": "model_1",
        "answer": """{output_1}"""
    },
    {
        "model": "model_2",
        "answer": """{output_2}"""
    }
]
```

Now please rank the models by the quality of their answers, so that the model with rank 1 has the best output. Your ranking should be based primarily on which response **provides the most accurate, relevant, and comprehensive content in direct relation to the query**.
Focus on the core information and how well the query's intent is addressed, rather than on aspects like writing style, formatting, length, or politeness, unless these directly impact the clarity or usefulness of the core content requested by the query.
Your response MUST be a valid JSON object. This JSON object must contain **ONLY ONE** key named `"ranking_list"`. The value of the `"ranking_list"` key MUST be a list of two dictionaries, where each dictionary contains the model name and its rank.
The structure should be **exactly** as follows:

```
{
    "ranking_list": [
        {"model": "<model-name>", "rank": <model-rank>},
        {"model": "<model-name>", "rank": <model-rank>}
    ]
}
```

Do **not** include any other text, explanations, or additional keys in the JSON object. Please provide the ranking that a well-informed human, focusing on these content-centric criteria, would likely give.
<—im_end—>

L. Qualitative Examples of Reward Model Behavior

This section presents several examples illustrating common failure modes or biases observed in standard reward models, highlighting the types of spurious correlations they may learn.

Example 1: Formatting Bias (Plumbing Diagram)

Question

Construct plumbing diagrams for a two-story house (describe its core in natural language)

Chosen Response

Main floor plumbing diagram for a two-story house: Waste pipes from the sinks, toilets, and bathtubs on the main floor should all connect to a common pipe that runs to the main drain line in the basement. Similarly, the water supply lines from the faucets, showers, and other fixtures should all connect to a common supply pipe which then runs to the first-floor water heater.

Rejected Response

Plumbing diagram for a two-story house:
Ground floor: - Toilet - Bathroom sink - Shower - Dishwasher
Second floor: - Toilet - Bathroom sink - Shower

Remarks

The rejected response uses a list format (a spurious attribute), which might be preferred by a reward model biased towards formattedness, even though the chosen response provides a more descriptive answer aligned with the prompt's request for natural language description.

Example 2: Politeness Bias (Medical Diagnosis)

Question

Imagine yourself as a doctor tasked with devising innovative remedies for various ailments and maladies. Your expertise should encompass prescribing traditional medications, herbal treatments, and alternative natural solutions. Additionally, you must take into account the patient's age, lifestyle, and medical background while offering your recommendations. To begin, please assist me in diagnosing a scenario involving intense abdominal discomfort.

Chosen Response

Certainly! To diagnose a scenario involving intense abdominal discomfort, I would first conduct a thorough examination of the patient's symptoms and medical history. This would involve a physical examination of the abdomen, including a pelvic exam, an abdominal X-ray, and a colonoscopy.

If the patient has a history of abdominal pain and nausea, he may be suffering from a condition called gastroparesis, also known as ”silent gas pain.” This is a serious condition characterized by severe abdominal pain and discomfort that can be severe or constant. The discomfort can be accompanied by nausea, vomiting, and diarrhea. It is important to seek medical attention immediately if the patient experiences any of these symptoms. Once I've determined the cause of the abdominal discomfort, I would then prescribe traditional medications to alleviate any pain or discomfort. These medications may include pain relievers like ibuprofen or acetaminophen, as well as anti-inflammatory drugs like ibuprofen and naproxen. If a traditional medication is not effective, alternative natural treatments may be prescribed. These treatments may include acupuncture, herbal remedies, and home remedies that may help to increase blood flow to the abdomen and alleviate any pain.

It's important to note that these are just a few examples of possible causes of abdominal discomfort. If you have any other symptoms or medical conditions that may be contributing to your discomfort, it is important to consult with a healthcare professional for a proper diagnosis and treatment plan.

Rejected Response

Sure, I can help you with that. Based on the symptoms you've described, it's possible that the patient may be suffering from gastroenteritis, commonly known as stomach flu. However, it's essential to consider other potential causes such as food poisoning, gastritis, or appendicitis. To determine the root cause, I'd recommend a thorough medical evaluation, including blood tests and imaging studies. Once the diagnosis is confirmed, appropriate treatment can be prescribed, which may include medication, hydration, and rest. It's crucial to consult a healthcare professional for accurate diagnosis and treatment.

Remarks

The rejected response begins with polite phrases (”Sure, I can help you with that.”), a potentially spurious attribute that reward models might learn to favor, even if the chosen response provides a more detailed diagnostic thought process aligned with the persona requested.

Example 3: Formatting and Instruction Following (Survey Question)

Question

Create a survey question that can help capture insights about the customer experience.

Chosen Response

How likely would you be to recommend our product or service to a friend or colleague?

Rejected Response

Sure! Here's a sample feedback form that can gather information about the user interface experience:

1. Overall, how would you rate your experience with the user interface? (Scale of 1-5, with 1 being poor and 5 being excellent)
2. Which specific features of the user interface did you find most useful or intuitive? (Open-ended text response)
3. Were there any features of the user interface that you found confusing or difficult to use? If yes, please provide details. (Open-ended text response)
4. Did you encounter any technical issues or bugs while using the user interface? If yes, please describe the issue(s) you faced. (Open-ended text response)
5. How would you rate the visual design and aesthetics of the user interface? (Scale of 1-5, with 1 being poor and 5 being excellent)
6. Were there any specific aspects of the visual design that you particularly liked or disliked? (Open-ended text response)
7. Did you find the user interface responsive and fast? (Yes/No)
8. Were there any specific areas where you experienced slow loading times or delays? If yes, please specify. (Open-ended text response)
9. How would you rate the organization and layout of the user interface? (Scale of 1-5, with 1 being poor and 5 being excellent)
10. Were there any specific areas where you felt the organization or layout could be improved? If yes, please provide details. (Open-ended text response)
11. Did you find the user interface easy to navigate and find what you were looking for? (Yes/No)
12. Were there any specific areas where you had difficulty navigating or finding information? If yes, please specify. (Open-ended text response)
13. How likely are you to recommend the user interface to others? (Scale of 1-5, with 1 being not likely at all and 5 being very likely)
14. Do you have any additional comments or suggestions for improving the user interface? (Open-ended text response)

Thank you for taking the time to provide your feedback!

Remarks

The rejected response fails to follow the instruction (provide a single question) and instead provides a lengthy form. A reward model might erroneously prefer it due to spurious features like length, formatting (numbered list), and politeness ("Sure!", "Thank you!").

Example 4: Formatting Bias vs. Content Quality (Electrician)

Question

What is the fastest way to become an electrician in the us

Chosen Response

Obtain a high school diploma or GED. Complete an electrical apprenticeship program, which typically lasts 4–5 years and combines on-the-job training with classroom instruction. Pass the Journeyman Electrician Exam to become licensed. Note that requirements and the time frame may vary depending on the state you are in.

Rejected Response

To become an electrician in the US, the fastest path typically involves the following steps: Foundational Knowledge: Formal Education: Consider pursuing a degree in electrical engineering or a related field. This will provide you with structured courses that cover the basics of electrical theory, circuit analysis, electronics, power systems, electromagnetism, and more. Online Courses and Tutorials: Websites like Coursera, edX, Udacity, and Khan Academy offer courses in electrical engineering and electronics. They can serve as an excellent supplement or alternative to traditional education. Hands-on Experience: DIY Projects: Start with small projects like building basic circuits, then gradually move to more complex ones. Kits are available online to help beginners learn the basics of circuit design and assembly. Internships and Apprenticeships: These provide invaluable real-world experience. Working under a mentor or experienced electrician can greatly enhance your learning. Certifications and Specializations: While not always mandatory, obtaining certifications can validate your expertise and open doors for specialized roles or areas within the field of electrics. Continuous Learning: Workshops and Seminars: Regularly attend these to stay updated with the latest techniques, tools, and best practices in the field. Books and Publications: Regularly read books, journals, and other publications related to electrical engineering and electronics. Networking: Professional Associations: Joining organizations like the Institute of Electrical and Electronics Engineers (IEEE) can provide networking opportunities, resources, and continued education. Online Forums and Groups: Websites like Stack Exchange (for electrical engineering), online forums, and even Reddit have active communities where you can ask questions, share knowledge, and learn from others. Practice and Experimentation: Regularly challenge yourself with new projects or problems to solve. The more you apply your knowledge, the more intuitive and deeply ingrained it will become. Stay Updated: Industry Trends: Electrics, like all fields of technology, is continuously evolving. Ensure you're aware of the latest tools, technologies, and methodologies in the field. Regular Refreshers: Periodically revisit foundational concepts to ensure your foundational knowledge remains strong. Soft Skills: While technical knowledge is crucial, don't underestimate the importance of communication, teamwork, problem-solving, and critical thinking skills. These are often essential, especially if you aim to take on leadership or interdisciplinary roles. Teach and Mentor: Once you've gained a certain level of expertise, consider teaching or mentoring others. This not only benefits the community but can also reinforce and expand your understanding. Becoming an electrician requires dedication, passion, and a willingness to continually push your boundaries. With focused effort and the right resources, you can achieve a high level of expertise and enjoy a fulfilling career in the world of electronics.

Remarks

The rejected response, while containing much unnecessary and potentially irrelevant information (e.g., suggesting an electrical engineering degree), uses extensive formatting (numbered lists with sub-points). A reward model overly sensitive to formattedness might prefer it over the more concise and directly relevant chosen response.