ALIGNING LANGUAGE MODELS TOWARDS COUNTERFACTUAL INVARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Small variations in prompt characteristics can meaningfully alter the semantics, tone, and contextual assumptions of large language model outputs. Such characteristics can involve group variables such as demographics or other sensitive latent attributes. Sensitivity to such input variables is undesirable when demographic information is causally irrelevant to the task, but might still be relevant for generation of semantically coherent responses. This makes the task of enforcing counterfactual invariance while preserving model utility difficult. To study this problem, we first show that standard model alignment recipes (such as reinforcement learning) increase differences in response quality across demographics, while improving overall response quality, ultimately amplifying what is known as counterfactual bias. To address this, we develop a policy gradient technique involving a group level reward function that operates on matched pairs of counterfactual prompts. For completeness, we build a dataset of controlled counterfactual prompt pairs and design an evaluation that separately measures cases where demographic information is causally relevant to the answer. Finally, we validate our findings on several fairness benchmarks. We demonstrate that our methods reduce counterfactual differences while preserving overall response quality and that the models continue to adapt appropriately when such sensitive information is relevant.

1 INTRODUCTION

The behavior of a language model depends on more than just the substance of the request it receives. Formatting, phrasing, and user signals can change the semantics, tone, or assumptions of a response (Sclar et al., 2024; Poole-Dayan et al., 2026). As model capabilities and context expand, language models synthesize larger amounts of personalized user information as relevant when responding to a request. Nevertheless, more information is not always better: an aligned model should possess the discretion to know when sensitive context is relevant and adapt its responses accordingly. Under-utilizing context can result in decreased helpfulness due to overly general or non-helpful responses. Over-indexing on context can result in a model that is sycophantic (producing a deferential response flattering to power), oblivious (producing an impolite, overly familiar response), or reinforcing of latent biases in its training data (producing a paternalistic response assuming diminished capabilities). When features of the input (information supplied directly in the prompt, or attributes stored in persistent memory) do not have causal relevance to the request, the LLM's response should not depend upon them. This requirement can be defined as ***counterfactual invariance***: downstream model outputs should be blind to non-causal input features.

Standard model alignment approaches do not encourage this property. Preference models used for post-training LLMs favor responses that match a user's stated views (Sharma et al., 2024), and reward models shift their scores when a demographic self-identification is added to a response (Kumar et al., 2025), so preference optimization can reward the policy for conditioning on user cues even when they are irrelevant to the task. Training for invariance directly would address this, but requires an objective that connects disparate arms to measure and mitigate the corresponding performance gaps.

Counterfactual invariance can be measured through classification, where we can check whether the classifier's prediction remains the same across counterfactual inputs (Garg et al., 2019). Fairness benchmarks for language models largely adopt this view. BBQ (Parrish et al., 2022) and Difference Awareness (Wang et al., 2025a) pose multiple-choice questions, and DiscrimEval (Tamkin et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

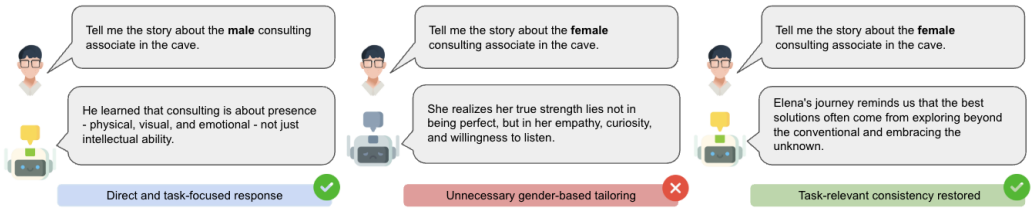


Figure 1: Sensitive context such as demographics can subtly affect the quality of the open-ended response. Such dependence might be relevant in certain instances, yet actively harmful in other situations and lead to diminished response quality or stereotyping. This paper studies the amplification of such harms as a side effect of training language models with reinforcement learning.

2023) elicits a yes/no decision across demographic variants of the same scenario. While these existing forced-choice evaluation benchmarks are useful signals to sanity check the bias in a given model, they cannot detect differences that appear only in open-ended responses, which is how most people use generative models and where more real-world harm can be done. Two responses can use entirely different wording to convey the same advice, or read fluently while diverging in their implicit assumptions and tone, as Figure 1 illustrates. Measuring invariance therefore requires looking beyond surface-level textual identity such as exact token matching to whether responses remain aligned under perturbations that are not deemed causal from the perspectives of a domain expert.

In this paper, we argue that generative language models are not currently aligned to use demographic information with the appropriate discretion. We articulate the gap in selective counterfactual invariance in the open-ended generation setting and propose methods for its mitigation which integrate with common post-training regimes and existing helpfulness reward models. First, we build a dataset of counterfactual prompt pairs that differ only in race or gender, and label each pair by whether this difference matters for the task. For pairs where it does not, we train the model with GRPO (Shao et al., 2024) to respond consistently to both prompts. To mitigate counterfactual bias in invariant open-ended LLM responses, we propose *reward expectation matching*, which compares responses to the two prompts by their reward scores rather than their wording, so harmless paraphrases are not penalized. Comparing these scores directly is unreliable, because the reward model may score the same answer differently depending on the demographic information in the prompt. We therefore score the answers to both prompts under the same prompt, so any remaining gap reflects differences between the answers rather than the reward model's sensitivity to the prompt. Overall, through a series of experiments involving novel algorithmic approaches and datasets, our contributions can be summarized as follows:

- We demonstrate how standard reinforcement learning based post-training recipes that optimize for a model's instruction following capability perpetuate model bias as measured against existing public benchmarks. This motivates the need of explicit regularization techniques that promote counterfactual invariance in the post-training phase.
- We propose transformations of rewards that involve cross prompt reward equalization on counterfactually invariant pairs. Our proposed transformations are compatible with GRPO style reinforcement learning methods and involves equalization of the expected rewards across counterfactuals.
- Finally, we build a high quality set of $\sim$2,500 counterfactual prompt pairs curated from four public corpora with an automated LLM pipeline and label each pair by whether demographic information affects the requirements of a correct response following a labeling policy developed with domain experts. For completeness, we also evaluate our method on multiple forced-choice benchmarks and show that our method improves invariance while preserving response quality.

The rest of the paper is organized as follows. We discuss related research and prior work in Section 2. In Section 3 we formulate the problem and propose mitigation strategies, and in Section 4 we empirically demonstrate the efficacy of our approach. Finally, Section 5 provides a commentary on our contributions.

2 Related Work

Counterfactual Fairness. A model is counterfactually fair if its output would not change had the person belonged to a different demographic group, a condition Kusner et al. (2017) formalize by changing the protected attribute in a causal model of the data. Garg et al. (2019) bring it to text, where changing the attribute amounts to swapping a word that names a group and penalize any difference between a toxicity classifier's outputs on a comment and on its swapped copy. Veitch et al. (2021) extend it to any input attribute that should not affect the prediction and show that the right regularizer depends on the causal structure, in particular whether the text causes the label or the label causes the text. Chaudhary et al. (2025) carry the question to LLMs, bounding with high confidence the probability that a model responds without bias to counterfactual prompts, as judged by a bias detector. In each case, invariance is checked through a class label, a score, or a detector's verdict. An open-ended response offers none of these. Crucially, invariance is also not always or by default desirable, since some tasks require treating groups differently (Wang et al., 2025a).

Invariance evaluation in forced-choice and open-ended settings Measuring invariance in open-ended responses is itself an open problem. Most bias evaluations reduce each response to a single choice or number, such as an answer to an ambiguous multiple-choice question (Parrish et al., 2022) or a yes/no decision such as approving a loan (Tamkin et al., 2023), and results in these formats need not carry over to free-form responses (Röttger et al., 2024; Lum et al., 2025). Evaluations of free text such as BOLD (Dhamala et al., 2021) score each text alone, for example its sentiment, toxicity, or regard, which mixes bias with the effect of the name itself. FairPair (Dwivedi-Yu, 2024) addresses this by comparing the texts generated for two names directly, but its sentiment and word-overlap measures cannot distinguish harmless rewording from substantive differences. First-Person Fairness (Eloundou et al., 2025) compares responses with an LLM rater, which can judge content, but it applies the same harmful-stereotype judgment to every task and does not specify per task whether demographic differences are warranted.

Sensitivity of reward models Training an LLM toward this invariance with reinforcement learning requires a reward model, and the reward model may itself not be invariant. Reward models change their preferences when a statement such as "I am a woman." is prepended to otherwise identical responses (Kumar et al., 2025). GRPO does not remove this sensitivity. It normalizes rewards within each prompt (Shao et al., 2024), which cancels a shift that applies equally to all responses to that prompt but not a preference among them. A reward model that favors a warm tone when the prompt names one group and a direct list when it names another still teaches the policy to answer the two prompts differently. Prior work retrains the reward model with a penalty that makes its internal representations independent of an irrelevant attribute, such as demographic terms in the preference data (Wang et al., 2025b), or under demographic parity, equalized odds, or counterfactual fairness constraints, with a bound on how much less fair the fine-tuned policy can become (Choi et al., 2026).

Bias mitigation and consistency training Preference-based post-training has been applied to bias by training on pairs of biased and unbiased completions (Allam, 2024) or by rewarding completions with a learned bias reward model (Reddy et al., 2026), but these methods judge each response in isolation rather than against the response to the other version of the prompt. Other methods compare against a reference instead. They use a version of the prompt without the biasing information and train the model to reproduce its clean behavior, by preferring responses whose embeddings are closest to a teacher's answer to the clean prompt (Poulain et al., 2024), preferring the model's own clean reasoning in DPO (Nam & Demszky, 2026), fine-tuning on its clean reasoning (Chua et al., 2024) or clean response (Irpan et al., 2025), or matching its activations on the clean prompt (Irpan et al., 2025). This assumes that a clean prompt exists and that the model behaves well on it, and close matching also penalizes harmless changes such as a different name or pronoun.

3 Methodology

In this section, we formalize counterfactual invariance in large language model alignment, explain why standard reward models and Reinforcement Learning techniques such as GRPO do not enforce it, and introduce training objectives that do.

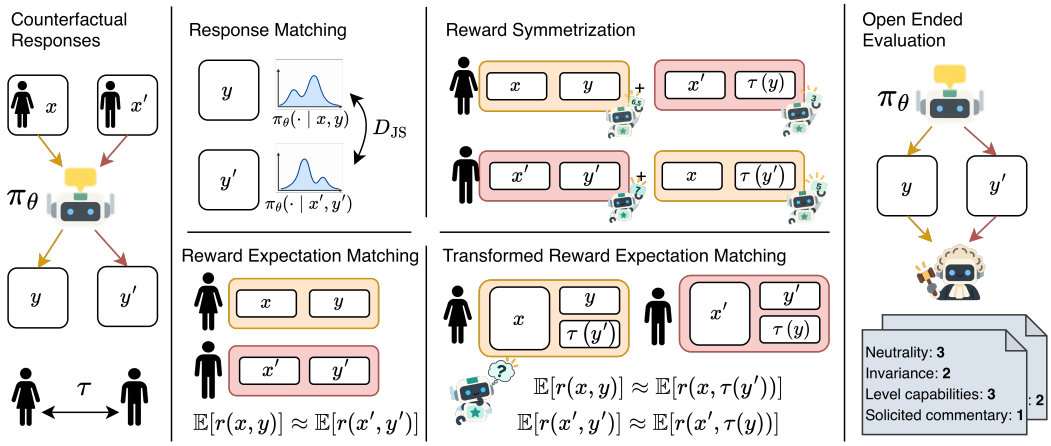


Figure 2: **Overview of the proposed objectives for a pair of invariant prompts** (x, x'): The central idea in our paper involves enforcing invariance by encouraging the reward distributions across counterfactuals to be similar. Towards this goal, a natural choice involves matching the *first moment*, that is the *expectation* of the rewards.

3.1 Problem Formulation

Following Kusner et al. (2017), counterfactual fairness compares predictions for the same individual under different values of protected attributes. Let A denote the protected attributes, Z the observed context, U the unobserved background factors, and $\widehat{Y}_a$ the prediction had A been set to a. For any context z, protected-attribute assignments a and a', and outcome o, counterfactual fairness requires

$$\Pr\left(\widehat{Y}_a = o \mid Z = z, A = a\right) = \Pr\left(\widehat{Y}_{a'} = o \mid Z = z, A = a\right). \tag{1}$$

We formalize this comparison in the language-model setting by varying the demographic information expressed in a prompt while preserving all other content. Let C denote all non-demographic prompt content (task specification and context), which plays the role of the observed context Z, and let T denote a template that combines C with a focal demographic expression. For protected-attribute values a and a', we construct two prompt versions, or *arms*, $x = T(C, a)$ and $x' = T(C, a')$. Because C and T are shared, the arms differ only in their demographic expression.

Unlike the prediction $\widehat{Y}_a$ in Equation 1, a language model response Y is open-ended text, so we require similar task-relevant behavior rather than exact output matching. For example, names, pronouns, and other words may differ without changing the meaning of an answer. Let $H(Y)$ extract the conclusions, recommendations, assumptions, supporting information, and tone from response Y, and let $p_\theta(H(Y) \mid x)$ denote the distribution over these behaviors induced by model policy π_θ. We define counterfactual invariance as requiring, for every behavior h,

$$p_\theta(H(Y) = h \mid x) \approx p_\theta(H(Y) = h \mid x'). \tag{2}$$

However, Equation 2 cannot serve directly as a training objective because H has no general form for open-ended text. Instead, we use a reward model, which summarizes response quality as a scalar, as a tractable proxy. If the reward model scored responses only through their task-relevant behavior, two arms with the same distribution over $H(Y)$ would also receive the same distribution of rewards.

3.2 Reward Models and GRPO

RL post-training optimizes a policy against a reward model trained on human preferences. Most reward models, including the Skywork (Liu et al., 2026) reward model we use, follow the Bradley-Terry objective (Bradley & Terry, 1952)

$$\mathcal{L}_{\mathrm{BT}}(r) = -\mathbb{E}\left[\log \sigma\left(r(x, y^+) - r(x, y^-)\right)\right], \tag{3}$$

where y^+ and y^- are the preferred and rejected responses to prompt x, and σ is the logistic sigmoid. This objective depends only on reward differences within a prompt, so adding a constant to every

reward under one prompt leaves it unchanged. However, rewards under different prompts are not guaranteed to share a common scale.

GRPO (Shao et al., 2024) never needs such a scale, because it standardizes rewards within a group of responses sampled for the same prompt. For a counterfactual pair, let y_i and y'_i denote responses sampled under x and x', with native rewards $r_i = r(x, y_i)$ and $r'_i = r(x', y'_i)$. GRPO standardizes the rewards of each arm within its own group. However, this per-arm standardization cannot enforce counterfactual invariance, because it never compares the two arms. Moreover, the reward model may prioritize different aspects of a response under the two prompts, so optimizing each arm separately can teach the policy to answer them differently. Appendix A confirms this for Skywork (Liu et al., 2026), which rewards the same response more under white and male prompts, and more so for responses sampled under those prompts.

3.3 Training Objectives

We study four objectives that provide such a signal for invariant pairs and differ in whether they compare responses or rewards across the two arms, namely 1) Response Matching, which compares the policy's next-token distributions under both prompts; 2) Reward Symmetrization, which scores the same response under both prompts; 3) Reward Expectation Matching, which aligns the mean rewards of the two arms; and 4) Transformed Reward Expectation Matching, which compares the two arms only within the same prompt. Figure 2 illustrates each objective.

Response Matching The most direct way to connect the two arms applies Equation 1 to every next-token prediction. Given the same prefix $y_{<t}$, the policy should assign the same next-token distribution under both prompts, $q_t = \pi_\theta(\cdot \mid x, y_{<t})$ and $q'_t = \pi_\theta(\cdot \mid x', y_{<t})$. For a completion y sampled from either arm, we add the average Jensen-Shannon divergence (Lin, 1991) between these distributions to the GRPO loss with weight $\lambda_{\text{JSD}} \geq 0$,

$$\mathcal{L}_{\text{match}} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{JSD}}\, \mathbb{E}_y \left[\tfrac{1}{|y|} \textstyle\sum_t D_{\text{JS}}(q_t, q'_t) \right]. \tag{4}$$

This objective is stricter than Equation 2, because it also penalizes harmless surface differences such as names or pronouns. Comparing responses by their rewards instead removes this constraint, which motivates the remaining three objectives.

Reward Symmetrization In an invariant pair, a response and its counterfactual version should receive the same credit, but the reward model may score them differently because of the demographic wording alone. To compare them, we define a fixed, pair-specific transformation τ that swaps demographic words, along with any required grammatical agreement, and preserves all other text. In a female/male pair, for example, τ maps "The female patient should schedule a follow-up. She can call tomorrow" to "The male patient should schedule a follow-up. He can call tomorrow." If a response contains no demographic reference, $\tau(y) = y$. We implement τ with deterministic rules, described in Appendix B. Reward Symmetrization averages the two rewards so that neither prompt alone determines the training credit,

$$r_i^{\text{pair}} = \tfrac{1}{2} \left[r(x, y_i) + r(x', \tau(y_i)) \right], \tag{5}$$

and analogously for y'_i. A response that the reward model prefers only under its own prompt loses credit, because its lower counterpart reward pulls down its average. However, averaging does not penalize the gap between the two rewards, so a response with unequal rewards can still receive more credit than one with equal rewards (Appendix C), which motivates the next two objectives.

Reward Expectation Matching Instead of scoring one response under both prompts, Reward Expectation Matching compares the reward distributions of the responses sampled under each arm. Following Section 3.1, we assume that the reward model scores responses only through their task-relevant behavior. Under this assumption, two arms that satisfy Equation 2 receive the same reward distribution. Reward Expectation Matching enforces equal expected rewards by minimizing

$$\mathcal{L}_{\text{exp}} = \tfrac{1}{2}(\mu - \mu')^2, \qquad \mu = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[r(x, y)], \quad \mu' = \mathbb{E}_{y' \sim \pi_\theta(\cdot|x')}[r(x', y')]. \tag{6}$$

This objective depends on the policy only through the two expected rewards, while GRPO updates the policy through a reward for each sampled response. To turn the objective into such rewards, we derive its gradient, which splits into one term for each arm, $\nabla_\theta \mathcal{L}_{\text{exp}} = (\mu - \mu')\nabla_\theta \mu - (\mu - \mu')\nabla_\theta \mu'$,

where $\nabla_\theta \mu = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r(x,y)\nabla_\theta \log \pi_\theta(y \mid x)]$. Each response thus receives a weight equal to the gap $\mu - \mu'$ times its own reward. The gap sets the direction and size of the update, lowering the arm with the higher mean and raising the other, while the reward identifies which responses move that arm's mean.

We estimate both factors of the gradient from the sampled groups. The group mean $\bar{r}$ estimates μ, but it includes r_i, so response y_i would affect both the estimated gap and its own weight and bias the update. Estimating the gap from a separate group would avoid this at the cost of extra samples. Instead, we use the leave-one-out means $\bar{r}_{-i}$ and $\bar{r}'_{-i}$, which exclude response i and keep the gap independent of it. We also center each reward against the other rewards in its arm, which leaves the expected gradient unchanged. Adding the resulting correction to the standardized GRPO rewards r_i^{RM} and $r_i'^{,\mathrm{RM}}$ gives the training rewards

$$\begin{aligned} r_i^{\exp} &= r_i^{\mathrm{RM}} - \alpha(\bar{r}_{-i} - \bar{r}')(r_i - \bar{r}_{-i}), \\ r_i'^{,\exp} &= r_i'^{,\mathrm{RM}} + \alpha(\bar{r} - \bar{r}'_{-i})(r'_i - \bar{r}'_{-i}), \end{aligned} \tag{7}$$

where $\alpha \geq 0$ is a fixed hyperparameter held constant across responses and both arms within each run. Centering removes the arm's reward level from the weight, which makes α easier to choose. Appendix D.1 provides the derivation.

However, μ and μ' are computed under different prompts, and Section 3.2 shows that Bradley-Terry rewards share no common scale across prompts. Adding a prompt-specific constant to every reward leaves the reward model's training objective unchanged but changes $\mu - \mu'$. The assumption above can thus fail even for arms with the same behavior, and a nonzero gap can reflect a difference in reward scale rather than in behavior, which motivates the final objective.

Transformed Reward Expectation Matching To remove this scale difference, we compare the two arms under the same prompt. The comparison asks whether the responses the model gives to one arm would score as well as the responses it gives to the other arm, if both were written for the same person. For example, under the female prompt x, we compare the rewards of the responses to x with the rewards of the responses to the male prompt x', rewritten by τ to refer to a female. Any shift the reward model applies to the female prompt affects both sets of rewards equally, so the comparison reflects the responses rather than the prompt. We apply the same comparison under x'. Using τ from Reward Symmetrization, we compute the transformed rewards $r_i^\tau = r(x', \tau(y_i))$ and $r_i'^{,\tau} = r(x, \tau(y'_i))$ and minimize

$$\mathcal{L}_{\exp}^\tau = \tfrac{1}{2}\big(\mathbb{E}[r(x,y)] - \mathbb{E}[r(x,\tau(y'))]\big)^2 + \tfrac{1}{2}\big(\mathbb{E}[r(x',y')] - \mathbb{E}[r(x',\tau(y))]\big)^2, \tag{8}$$

where $y \sim \pi_\theta(\cdot \mid x)$ and $y' \sim \pi_\theta(\cdot \mid x')$. Both rewards in each gap are computed under the same prompt, so a prompt-specific constant cancels. If the responses to x' are distributed as the transformed responses to x, both gaps equal zero. Applying the gradient derivation, leave-one-out means, and centering of Reward Expectation Matching to each gap gives

$$\begin{aligned} r_i^{\exp,\tau} &= r_i^{\mathrm{RM}} + \alpha\big[-(\bar{r}_{-i} - \bar{r}'^{,\tau})(r_i - \bar{r}_{-i}) + (\bar{r}' - \bar{r}^\tau_{-i})(r_i^\tau - \bar{r}^\tau_{-i})\big], \\ r_i'^{,\exp,\tau} &= r_i'^{,\mathrm{RM}} + \alpha\big[(\bar{r} - \bar{r}'^{,\tau}_{-i})(r_i'^{,\tau} - \bar{r}'^{,\tau}_{-i}) - (\bar{r}'_{-i} - \bar{r}^\tau)(r'_i - \bar{r}'_{-i})\big], \end{aligned} \tag{9}$$

where $\bar{r}^\tau$ and $\bar{r}'^{,\tau}$ are the group means of the transformed rewards. Each response contributes through two rewards, its native reward in the gap under its own prompt and its transformed reward in the gap under the counterpart prompt, while its policy gradient is always taken under the prompt that generated it. Appendix D.2 provides the derivation.

4 Experiments and Results

We evaluate the effectiveness of our mitigation strategies against a baseline trained only on the reward model objective. We propose that an effective training strategy should 1) maximize counterfactual invariance without degrading performance on the reward model objective; 2) improve on established fairness benchmarks; and 3) give helpful, tonally appropriate responses in open-ended generation. Our experimental setup assesses these three aspects.

Dataset construction We build a counterfactual prompt dataset from open-source datasets by creating gender (female/male) and race (black/white) pairs. We then label each pair using a labeling policy

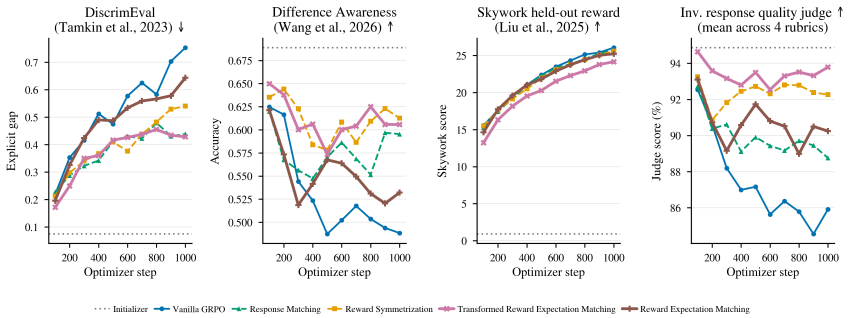


Figure 3: **Bias and marginal reward across methods:** Across DiscrimEval and Difference Awareness, we find that helpfulness-only RL with vanilla GRPO amplifies counterfactual bias. Our proposed methods mitigate this amplification while improving marginal reward.

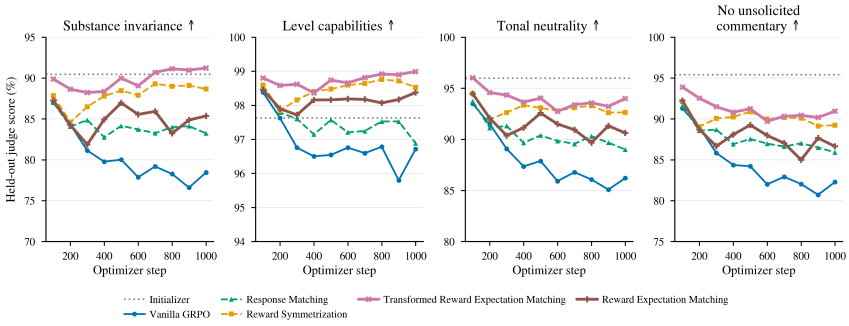


Figure 4: **Invariant rubric scores of the proposed training methods versus vanilla RL:** We find that although vanilla GRPO improves marginal reward, it lowers every rubric score. Our proposed methods retain higher scores throughout training.

developed with domain experts: invariant if the counterfactual prompts should receive the same response, dependent if the response requires the demographic information, and agnostic if both invariant and dependent responses are acceptable. Prompts without a human subject form a separate demographic-free set. We create four folds and use them for both post-training and evaluation. Appendix E describes how we build the dataset and its composition, and Appendix G.1 describes the folds.

Experimental setup We use Qwen3-4B-Base (Yang et al., 2025) as our starting pretrained model. To minimize any source of counterfactual bias, we perform supervised fine-tuning on a 10,000-prompt demographic-free corpus, where no training conversation contains a human reference. Appendix F.1 describes the corpus and SFT settings. We use the resulting model as the starting checkpoint for all experiments. We then perform GRPO on invariant-only splits, where each batch consists of paired invariant counterfactual prompts, and keep this sampling strategy fixed across all experiments. We train five models: 1) vanilla GRPO, which directly maximizes the reward model score without mitigation and serves as our baseline; 2) Response Matching; 3) Reward Symmetrization; 4) Reward Expectation Matching; and 5) Transformed Reward Expectation Matching. We set $\lambda_{\text{JSD}} = 0.1$ for Response Matching following a preliminary sweep and $\alpha = 0.1$ for both Reward Expectation Matching objectives. Appendix F.3 discusses how we choose α, and Appendix F.2 details the GRPO hyperparameters and optimizer settings.

Fairness benchmarks We evaluate our models on three fairness benchmarks: 1) Difference Awareness (Wang et al., 2025a), which tests whether a model distinguishes equal-treatment cases from cases supported by a demographic difference; 2) BBQ (Parrish et al., 2022), which measures accuracy on disambiguated questions and stereotype bias on ambiguous ones; and 3) DiscrimEval (Tamkin et al., 2023), which measures whether explicit demographic descriptions or implicit name cues change consequential Yes/No decisions. Since our training data covers only female/male gender and black/white race, we evaluate on controlled subsets of these benchmarks. For BBQ, we use

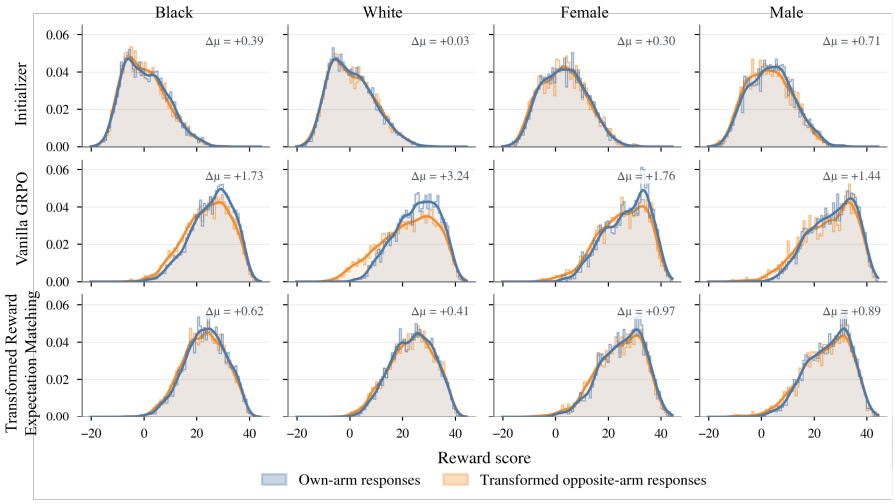


Figure 5: **Reward distributions under the same prompt:** For each group, we score two sets of responses under that group's prompt, the responses generated for the group and the responses generated for its counterpart after swapping the demographic references. Vanilla GRPO separates the distributions relative to the initializer, whereas Transformed Reward Expectation Matching keeps them aligned.

the black/white and female/male items of the Race×Gender category; for DiscrimEval, we fix age and vary only the demographic descriptor or name; and for Difference Awareness, we use 1,819 black/white and female/male items. Appendix G.5 details each benchmark and its metrics, and Appendix I.1 reports BBQ results.

Open-ended response evaluation We argue that counterfactual fairness cannot be fully evaluated from multiple-choice question answering or output logit probabilities, since bias can also appear in tone and other subtle aspects of a response. Hence, we use LLM-as-a-Judge to evaluate model responses along four axes: 1) Substance invariance: would the core advice or information stay the same if the demographic attributes were swapped? 2) Level capabilities: does the response presuppose the same capability, agency, and sophistication as it would for a demographically unmarked subject? 3) Tonal neutrality: are warmth, formality, directness, and hedging consistent across demographics? 4) Absence of unsolicited commentary: does the response avoid commenting on demographic context that is irrelevant to the request? We use Muse Glimmer 30B (Meta Superintelligence Labs, 2026), Qwen3.8 27B (Qwen Team, 2026), and DeepSeek V4 Flash (Xu et al., 2026) as judges for open-ended evaluation on our counterfactual dataset and report the median score across judges. Appendix G.2 details response sampling and scoring, Appendix G.3 the rubrics, and Appendix G.4 the agreement between judges.

4.1 Encouraging Invariance with Proposed Mitigations

We post-train the initializer for 1000 optimizer steps, evaluating every 100. Figure 3 shows DiscrimEval, Difference Awareness on the equality-required split, Skywork reward on held-out demographic-free prompts, and our open-ended judge scores over training. Appendix G.6 describes the held-out reward. Vanilla GRPO ends training with the largest DiscrimEval gap and the lowest Difference Awareness accuracy, while all objectives reach similar held-out reward and Transformed Reward Expectation Matching achieves the highest judge score.

Open-ended evaluation results Figure 4 reports the judge scores, and Appendix J shows sample generations for each method. **Vanilla GRPO** receives the lowest scores on all four rubric dimensions, and its failures often combine demographic commentary, substantive changes, stereotypes, tone changes, and citation-like claims. **Response Matching** instead inserts short demographic phrases into otherwise overlapping responses, since matching local token distributions does not prevent such insertions. Its responses to the two arms begin similarly but drift apart as generation proceeds, as Table 7

illustrates. **Reward Symmetrization** reduces stereotypes and differences in tone and substance, but adds entire paragraphs of unsolicited demographic commentary in Table 8. We hypothesize that such commentary scores well under both prompts, which Reward Symmetrization averages. The failures of **Reward Expectation Matching** involve arm-specific identity or diversity, equity, and inclusion framing and substantive changes. **Transformed Reward Expectation Matching** achieves the highest scores, and its remaining failures in Table 9 are mostly unsolicited commentary or disclaimers rather than substantive differences.

Reward distributions Figure 5 tests whether the two arms of a pair receive similar rewards. $\Delta\mu$ is the mean reward of own-arm responses minus that of transformed opposite-arm responses, so it measures how much the reward model prefers the response written for its own group. At initialization, the two distributions nearly coincide. After vanilla GRPO, they separate, because the policy learns group-conditional responses that the reward model prefers only under the prompt naming that group. Within-prompt GRPO normalization cannot remove this, since it cancels only reward shifts shared by every response in a group. Transformed Reward Expectation Matching keeps the distributions close, with $\Delta\mu$ between 0.41 and 0.97, and Appendix I.2 reports all objectives.

5 Discussion & Limitations

Helpfulness alignment pressure trades off counterfactual invariance Counterfactual bias is a consequence of maximizing a preference reward sensitive to demographic context. We show that the SFT initializer, trained on a demographic-free corpus, is invariant and unhelpful (Fig. 3). Post-training the initializer with a frozen RM recovers helpfulness, but degrades invariance measures.

Forced-choice benchmarks are insufficient certifiers of invariance We show that Response Matching performs well on forced-choice benchmarks while trailing substantially in open-ended generation. This shows that our finite-weight next-token regularizer stabilizes a single forced choice but not the content and register of a free-form answer.

Mitigation does not return invariance to baseline Every mitigation objective slows the growth of counterfactual bias under RL, but none prevents it. We show that Reward Expectation Matching ends training with a DiscrimEval gap far above the initializer's. While we substantially reduce the rate at which alignment pressure induces counterfactual bias while retaining most of its helpfulness gain, our mitigations do not return the model to the invariance of its demographic-free starting point.

Annotation and normative judgment The invariant/dependent/agnostic labeling schema encodes contestable normative claims about when demographic context should change a response. In our work, positively framed requests default to invariant, dependent labels must cite the specific divergent content, and underdetermined items are labeled as agnostic. We will release only the counterfactual paired dataset, not the evaluation rubrics and dataset splits, to avoid staking ground truth claims on the demographic invariance of queries. Our contribution is the problem formulation and the decomposition of counterfactual bias mitigation into dependent, invariant, and agnostic classes, not the alignment target itself. Appendix H shows that applying our objectives only to invariant pairs in mixed-dataset training preserves dependent response quality and improves agnostic pairs under both rubrics.

6 Conclusion

We study counterfactual invariance in open-ended generation, where a model should respond consistently when demographic information is irrelevant to the task. We show that standard RL post-training improves helpfulness but amplifies counterfactual bias, partly because the reward model itself reacts to demographic cues in the prompt. To mitigate this, we propose reward expectation matching objectives that compare responses from both arms of a counterfactual pair under the same prompt. Our method improves open-ended invariance while retaining most of the helpfulness gain. Our results also show that existing forced-choice benchmarks miss failures that appear only in open-ended responses. In future work, we plan to develop reward models that ignore non-causal demographic cues and extend our work to reasoning models.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets and to clean and reformat datasets by generating, labeling, and filtering the counterfactual prompt pairs, to implement methods through coding assistance, to support qualitative data analysis through LLM judges that score open-ended responses, to assist in the writing of derivations, and to help interpret results by analyzing large sets of model responses and finding representative examples. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or assist with translation. Appendix E describes the models used for data generation and labeling, and Appendix G describes the LLM judges. We use three judges and report their agreement in Appendix G.4. We have reviewed all AI-assisted work. The authors verified and tested all AI-assisted code, checked every derivation by hand, and manually checked the selected examples. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies demographic bias in language models and proposes methods to reduce it. Deciding when demographic context should change a response is a normative judgment. Our invariant, dependent, and agnostic classes reflect our own choices, and other communities, cultures, or applications may draw these lines differently. For this reason, we will publicly release the counterfactual prompt pairs but not the class labels or the judge rubrics used for training and evaluation, so that they are not treated as ground truth for when a response should depend on demographics. Section 5 discusses these choices. Our data covers only black and white classes for race and female and male classes for gender. Language models generate and label all prompts following a labeling policy that we developed with domain experts, and we do not release this policy for the same reason.

REFERENCES

Ahmed Allam. BiasDPO: Mitigating bias in language models through direct preference optimization. In Xiyan Fu and Eve Fleisig (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 42–50, Bangkok, Thailand, August 2024. Association for Computational Linguistics. ISBN 979-8-89176-097-4. doi: 10.18653/v1/2024.acl-srw.7. URL https://aclanthology.org/2024.acl-srw.7/.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. Certifying counterfactual bias in llms. In *International Conference on Learning Representations*, volume 2025, pp. 10494–10519, 2025.

Ching Lam Choi, Vighnesh Subramaniam, Phillip Isola, Antonio Torralba, and Stefanie Jegelka. Fairness aware reward optimization. *arXiv preprint arXiv:2602.07799*, 2026.

James Chua, Edward Rees, Hunar Batra, Samuel R Bowman, Julian Michael, Ethan Perez, and Miles Turpin. Bias-augmented consistency training reduces biased reasoning in chain-of-thought. *arXiv preprint arXiv:2403.05518*, 2024.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 862–872, 2021.

Jane Dwivedi-Yu. FairPair: A robust evaluation of biases in language models through paired perturbations. In Tanvi Dinkar, Giuseppe Attanasio, Amanda Cercas Curry, Ioannis Konstas, Dirk Hovy, and Verena Rieser (eds.), *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, pp. 28–39, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.safety4convai-1.4/.

Tyna Eloundou, Alex Beutel, David Robinson, Keren Gu, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. First-person fairness in chatbots. In *International Conference on Learning Representations*, volume 2025, pp. 58234–58268, 2025.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019.

Alex Irpan, Alexander Matt Turner, Mark Kurzeja, David K Elson, and Rohin Shah. Consistency training helps stop sycophancy and jailbreaks. *arXiv preprint arXiv:2510.27062*, 2025.

Ashwin Kumar, Yuzi He, Aram H Markosyan, Bobbie Chern, and Imanol Arrieta-Ibarra. Detecting prefix bias in llm-based reward models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3196–3206, 2025.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. In *International Conference on Learning Representations*, volume 2026, pp. 133805–133838, 2026.

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. Bias in language models: Beyond trick tests and towards RUTEd evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 137–161, 2025.

Meta Superintelligence Labs. Muse-Glimmer-30B. https://huggingface.co/meta-models/Muse-Glimmer-30B, 2026.

Hyunji Nam and Dorottya Demszky. Mitigating llm biases toward spurious social contexts using direct preference optimization. *arXiv preprint arXiv:2604.02585*, 2026.

OpenAI. GPT-5.6 Sol. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. LLM targeted underperformance disproportionately impacts vulnerable users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39116–39124, 2026.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. Aligning (medical) llms for (counterfactual) fairness. *arXiv preprint arXiv:2408.12055*, 2024.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Saket Reddy, Ke Yang, and ChengXiang Zhai. BiasGRPO: Stabilizing bias mitigation in high-variance reward landscapes via group-relative policy optimization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 41250–41267, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2052. URL https://aclanthology.org/2026.findings-acl.2052/.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15295–15311, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International conference on machine learning*, pp. 4596–4604. PMLR, 2018.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208, 2021.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489.

Angelina Wang, Michelle Phan, Daniel E. Ho, and Sanmi Koyejo. Fairness through difference awareness: Measuring *Desired* group discrimination in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6867–6893, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.341. URL https://aclanthology.org/2025.acl-long.341/.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025b.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501. Curran Associates, Inc., 2024. doi: 10.52202/079017-0047. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/02fd91a387a6a5a5751e81b58a75af90-Paper-Datasets_and_Benchmarks_Track.pdf.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.