

ARF-RLHF: Adaptive Reward-Following for RLHF through Emotion-Driven Self-Supervision and Trace-Biased Dynamic Optimization

Yuxuan Zhang

South China Normal University

University of Aberdeen

Avenum_Z@outlook.com

Abstract

Current RLHF methods such as PPO and DPO typically reduce human preferences to binary labels, which are costly to obtain and too coarse to reflect individual variation. We observe that expressions of satisfaction and dissatisfaction follow stable linguistic patterns across users, indicating that more informative supervisory signals can be extracted from free-form feedback. Building on this insight, we introduce Adaptive Reward-Following (ARF), which converts natural feedback into continuous preference trajectories and optimizes them using the novel TraceBias algorithm. Across diverse LLMs and preference domains, ARF consistently outperforms PPO and DPO, improving alignment by up to 7.6%. Our results demonstrate that continuous reward modeling provides a scalable path toward personalized and theoretically grounded RLHF.

1 Introduction

The rapid evolution of large language models (LLMs) has shifted the challenge of alignment from factual accuracy to deeper personalization—adapting model behavior to the intent and preferences of individual users. Reinforcement Learning from Human Feedback (RLHF) is now the de facto fine-tuning strategy, powering models such as GPT-4 (OpenAI et al., 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025a), and Llama-3 (Grattafiori et al., 2024). However, prevailing RLHF methods such as PPO (Schulman et al., 2017b) and DPO (Rafailov et al., 2024) depend on large-scale binary preference annotations. These signals are expensive to collect, coarse in resolution, and biased toward group-level tendencies rather than individual needs.

Recent efforts to reduce annotation burden—such as RLAIIF (Lee et al., 2024), crowd-sourced preference distillation (Zhang et al., 2024),

and KTO (Ethayarajh et al., 2024)—have improved sample efficiency but still rely on externally sourced annotations or handcrafted prompt pipelines. Without rethinking how feedback itself is represented, such approaches cannot capture the evolving trajectory of individual preferences and will continue to inherit population-level biases.

To address this gap, we draw inspiration from psycholinguistic research. *Common Ground Theory* (Clark, 1996) shows that people share conventions for signaling satisfaction or dissatisfaction; LIWC analyses (Tausczik and Pennebaker, 2010) confirm stable emotion-word usage across cultures; and *Emotion Prototype Theory* (Fehr and Russell, 1984) explains how affect can be inferred from prototypical linguistic cues. Together, these findings suggest that free-form feedback naturally encodes graded satisfaction signals. We leverage these regularities not as fixed templates, but as priors for interpreting a single user’s feedback in context—transforming open-ended responses into a posterior estimate of satisfaction.

Guided by this insight, we propose the **Adaptive Reward-Following (ARF)** framework, a self-supervised RLHF pipeline designed to minimize manual labeling while capturing fine-grained, per-user preferences. As illustrated in Figure 1, ARF:

1. infer posterior satisfaction signals from natural user feedback via a *Static Satisfaction Scorer*;
2. augment samples through synonym substitution, truncation, and preference-weighted rebalancing;
3. train a lightweight *ARF Scorer* with soft labels to directly predict continuous satisfaction scores;
4. fine-tune the LLM with the proposed *Trace-Bias* algorithm, which optimizes reward trajectories rather than binary labels.

This workflow removes the reliance on costly pairwise annotations, while enabling precise mod-

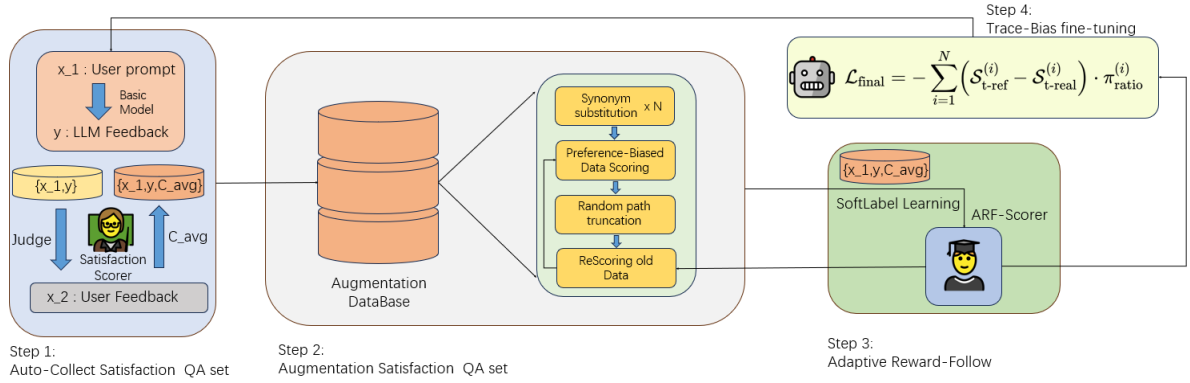


Figure 1: Illustrating the overall workflow of our framework. We begin by deriving posterior satisfaction estimates from natural user feedback via a Static Satisfaction Scorer (Step 1). These samples are then stored and augmented through synonym substitution, truncation, and reweighting to form a diversified reward corpus (Step 2). The ARF scorer is trained with soft labels to predict satisfaction scores and is continuously updated (Step 3). Finally, the TraceBias algorithm leverages ARF-generated rewards to fine-tune the LLM (Step 4), completing a fully self-supervised RLHF pipeline.

eling of evolving, user-specific preferences. Experiments on Qwen-2/2.5 (Qwen et al., 2025), Gemma-2 (Team et al., 2024), and Llama-3.2 (Grattafiori et al., 2024) across four preference domains show that ARF not only matches strong baselines but outperforms PPO by 3.3% and DPO by 7.6%.

Contributions. Our work makes three key contributions:

- We introduce ARF, the first self-supervised RLHF framework that learns continuous satisfaction trajectories directly from free-form user feedback.
- We propose *TraceBias*, an RLHF fine-tuning algorithm that leverages these trajectories to optimize alignment without binary comparisons. We also prove its theoretical consistency with standard RLHF methods (see Appendix B).
- We demonstrate across multiple LLM families and domains that ARF achieves scalable and personalized alignment, reaching or surpassing the performance of several baselines while reducing annotation cost.

Together, these results position ARF as a step toward next generation RLHF: alignment that is scalable, personalized, and grounded in both linguistic theory and empirical performance.

2 Background

2.1 RLHF as Preference-Based Optimization

Reinforcement Learning from Human Feedback (RLHF) has become the dominant paradigm for aligning large language models (LLMs) with hu-

man intent. Despite implementation variations, most pipelines share a constant foundation: (i) preference modeling via the Bradley–Terry (BT) framework (BRADLEY, 1955), and (ii) policy optimization using variants of policy gradient.

Pipeline. RLHF typically proceeds in three stages. First, a model is pretrained on large-scale corpora and optionally finetuned (SFT) on instruction data. Second, a preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$ is collected, where $y_w \succ y_l$ indicates that humans prefer y_w over y_l for input x . Third, optimization is performed to maximize expected return under the inferred preference structure.

Preference Modeling. The BT model assumes the probability of preferring y_w to y_l is

$$\mathcal{P}(y_w \succ y_l | x) = \sigma(\mathcal{R}(x, y_w) - \mathcal{R}(x, y_l)), \quad (1)$$

where $\mathcal{R}(x, y)$ is a learned reward function and σ the logistic function.

Policy Optimization. Given a reward proxy, the policy π_θ is updated using standard policy gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t | s_t) A^\pi(s_t, a_t) \right], \quad (2)$$

with A^π the advantage function. In practice, proximal methods such as PPO (Schulman et al., 2017b) are used to stabilize training.

2.2 Two Dominant Formulations: PPO and DPO

Most existing work falls into two closely related approaches:

PPO with Reward Modeling. PPO first fits a reward model via BT loss:

$$\mathcal{L}_R = -\mathbb{E}_{(x, y_w, y_l) \sim D} \log \sigma(\mathcal{R}(x, y_w) - \mathcal{R}(x, y_l)), \quad (3)$$

and then optimizes the policy with clipped policy gradients.

DPO as a Direct Alternative. DPO (Rafailov et al., 2024) removes the explicit reward model, directly maximizing the margin between preferred and dispreferred responses:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim D} \log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right), \quad (4)$$

where π_{ref} is a reference model and β a temperature hyperparameter.

Limitations. While PPO/DPO-based fine-tuning methods and their variants differ in certain design choices, they are both grounded in the same actor-critic framework (see Appendix B for an in-depth discussion) and share common structural constraints: (1) reliance on discrete preference pairs, (2) dependence on costly human annotation, and (3) limited personalization, since alignment is defined at the population level. These limitations motivate approaches that move beyond binary comparisons and static reward proxies.

3 Methodology

ARF-RLHF builds on the framework in Section 2.1 and consists of three stages: model initialization, preference scoring, and policy optimization. For initialization, we use an open-source state-of-the-art LLM. Preference scoring relies on a composite system that combines a high-precision satisfaction scorer with an adaptive ARF scorer that evolves over time. The evolving signals guide policy optimization through the TraceBias algorithm. TraceBias fine-tunes the model with soft reward signals from the scorers, removing the need for binary labels and allowing more nuanced updates.

The following subsections elaborate on each component. We begin with the scoring system, describing the interaction between the static scorer and the adaptive preference tracker. Next, we present the score-shift mechanism within the augmentation database, which includes temporal bias correction and dual-scorer alignment. Finally, we analyze the TraceBias algorithm and its role in optimizing policies within the ARF-RLHF pipeline.

3.1 Adaptive Reward-Following (ARF) Scorer

Studies show that human communication conveys explicit meaning while also signaling satisfaction and willingness to continue (Chen and Chen, 2016; Shanahan et al., 2006; Henry et al., 2021; Prabhunoye et al., 2017). Based on this, we introduce two complementary scorers: a static satisfaction scorer that collects quality estimates, and an ARF scorer that updates periodically to track evolving preferences.

3.1.1 Static Satisfaction Scorer

Both scorers use the lightweight RoBERTa-mini¹ (Liu et al., 2019) architecture, which offers low latency and solid semantic understanding.

For self-supervised reward modeling, the static scorer predicts the quality of a (prompt, response) pair from the user’s follow-up reply. Specifically, it takes the follow-up as input and outputs a satisfaction score reflecting the sentiment toward the previous system response.

We project the final hidden states of RoBERTa-mini to three sentiment classes: *negative*, *neutral*, and *positive*, and aggregate token-level logits to obtain a sequence-level satisfaction distribution:

$$\mathcal{C}_3 = \text{Linear}(\mathcal{H}_{\text{Last}}) \quad (5)$$

$$\mathcal{C}_{\text{avg}} = \text{Softmax} \left(\frac{1}{L} \sum_{n=1}^L \mathcal{C}_3^{(n)} \right) \quad (6)$$

Here, $\mathcal{H}_{\text{Last}}$ denotes the final hidden states from RoBERTa-mini, and $\mathcal{C}_3 \in \mathbb{R}^{L \times 3}$ represents token-level satisfaction logits, where L is the input sequence length. The final prediction $\mathcal{C}_{\text{avg}} \in \mathbb{R}^3$ summarizes sequence-level satisfaction through mean pooling followed by softmax normalization. The importance of this three-class structure is further discussed in Discussion 5.2.

These static predictions are collected as soft labels to train the ARF scorer, which learns to assign reward scores to collected or new (prompt, response) pairs offline. The ARF scorer then serves as the reward function in TraceBias, guiding LLM fine-tuning without manual annotations.

3.1.2 ARF Scorer

The ARF scorer is fine-tuned during interactions to adapt to changing preferences. At each training

¹We use the standard RoBERTa-base model (Liu et al., 2019). The term ‘mini’ in the main text is used descriptively to highlight its compact size relative to large language models, not to denote a separate model variant.



Figure 2: We compare the gradient norm statistics of PPO, using a clip range $\epsilon = 0.2$ as in the original paper (Schulman et al., 2017b) and TraceBias with DAM. DAM exhibits lower variance and more stable gradient magnitudes, suggesting improved training stability and potential for enhanced performance. (V is shown in appendix I)

step, it outputs a predicted satisfaction distribution $\hat{\mathcal{C}}$ for LLM feedback. The prediction is supervised using soft labels derived from user feedback and historical interactions.

To speed up convergence while keeping baseline estimates stable, the ARF scorer builds on the static satisfaction scorer. The averaged satisfaction vector \mathcal{C}_{avg} , distilled from user follow-up messages, provides soft guidance for supervision. We employ the standard cross-entropy loss:

$$\mathcal{L}_{\text{supervised}} = \text{CE}(\hat{\mathcal{C}}, \mathcal{C}_{\text{avg}}) \quad (7)$$

To mitigate overfitting and catastrophic forgetting when real-time data is scarce, we use an Experience Replay (ER) mechanism. A sampling ratio ER_{ratio} switches training probabilistically between historical data and current feedback:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{ER}} = \text{CE}(\hat{\mathcal{C}}, \mathcal{C}_{\text{static}}), & \text{if } p < \text{ER}_{\text{ratio}} \\ \mathcal{L}_{\text{supervised}} = \text{CE}(\hat{\mathcal{C}}, \mathcal{C}_{\text{avg}}), & \text{otherwise} \end{cases}$$

Here, $\mathcal{C}_{\text{static}}$ denotes labels from the static satisfaction dataset (e.g. DailyDialog (Li et al., 2017), GoEmotions (Demszky et al., 2020)), and p is a random variable, sampled from a uniform distribution $p \sim \text{Uniform}(0, 1)$. This alternating strategy lets the ARF scorer benefit from both stable historical signals and dynamic user feedback, improving generalization and robustness.

3.2 Augmentation Database

For better use limited user feedback, we build an Augmentation Database that expands training data with synonym substitution and random trace truncation. We use a score-shift mechanism, blend-

ing static and adaptive distributions, ensuring augmented samples align with changing preferences.

3.2.1 Preference-Biased Fusion Mechanism

To ensure robust and adaptive scoring, we unify the preference-biased update into a single mechanism. Given two candidate distributions \mathcal{C}_A and \mathcal{C}_B , we compute a dynamic weighting coefficient:

$$\mathcal{S}_{\text{cos}} = \sigma((\text{CosSim}(\mathcal{C}_A, \mathcal{C}_B) - 0.5) \cdot S_{\text{sig}}), \quad (8)$$

where σ denotes the sigmoid, CosSim the cosine similarity, and S_{sig} a sensitivity scaling factor (Appendix D for selection). The fused score is then:

$$\mathcal{C}_{\text{fused}} = \mathcal{C}_A \cdot \mathcal{S}_{\text{cos}} + \mathcal{C}_B \cdot (1 - \mathcal{S}_{\text{cos}}) \quad (9)$$

This covers two cases:

- **Augmented Data Scoring:**

$$\mathcal{C}_A = \mathcal{C}_{\text{ARF}_{\text{avg}}}, \quad \mathcal{C}_B = \mathcal{C}_{\text{basic}_{\text{avg}}}$$

$\mathcal{C}_{\text{ARF}_{\text{avg}}}$: mean score from the adaptive ARF scorer.

$\mathcal{C}_{\text{basic}_{\text{avg}}}$: mean score from the static scorer.

- **Historical Re-Scoring:**

$$\mathcal{C}_A = \mathcal{C}_{\text{new}_{\text{avg}}}, \quad \mathcal{C}_B = \mathcal{C}_{\text{old}_{\text{avg}}}$$

$\mathcal{C}_{\text{new}_{\text{avg}}}$: mean score from the updated ARF scorer.

$\mathcal{C}_{\text{old}_{\text{avg}}}$: mean score stored from past evaluations.

This balances stability from static or historical references with adaptability to evolving preferences, preventing overfitting while maintaining continuity.

3.3 TraceBias Algorithm

Given enough preference-rich interaction data, we optimize policies with a score-based actor-critic algorithm called **TraceBias**. While theoretically

aligned with PPO and DPO, TraceBias uses a token-wise strategy and new normalization methods for stable, fine-grained optimization, without binary comparisons (see pseudocode in Appendix C).

The algorithm has two key components: (1) a **Double Average Method (DAM)** that normalizes satisfaction scores and token-level policy ratios, stabilizing training across variable-length sequences; and (2) an advantage function derived from trajectory-level score differences between fine-tuned and reference models, serving as the main optimization signal.

3.3.1 Double Average Method (DAM)

We address gradient inconsistencies that arise from variable sequence lengths. Rather than using explicit gradient clipping—which can remove informative signals—we normalize both satisfaction scores and token-level policy ratios.

Let T_i be the sequence length (number of tokens in an output), N the number of turns in a trajectory, and $i \in \{1, \dots, N\}$ the turn index. Denote $x^{(i)}$ as the input at turn i , and $y_t^{(i)}$ as its t -th generated token. P_θ and P_{old} are the current and reference policy probabilities. We define the average log-probabilities under each policy and their ratio as:

$$\pi_\theta^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_\theta(y_t^{(i)} | x^{(i)}) \quad (10)$$

$$\pi_{\text{old}}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_{\text{old}}(y_t^{(i)} | x^{(i)}) \quad (11)$$

$$\pi_{\text{ratio}}^{(i)} = \exp(\pi_\theta^{(i)} - \pi_{\text{old}}^{(i)}) \quad (12)$$

By combining satisfaction score normalization \mathcal{C}_{avg} (Eq. 6) with normalized token-level ratios, DAM reduces gradient imbalance—where long sequences otherwise dominate updates and short ones diminish. This produces more stable gradients than direct clipping, as illustrated in Fig. 2.

3.3.2 Trace Scores with Discounted Step-wise Evaluation

With token-level optimization stabilized, we compute score differences between generated and reference trajectories— $\mathcal{S}_{\text{t-real}}$ and $\mathcal{S}_{\text{t-ref}}$ —to estimate the advantage function.

At each evaluation step $j \in \{1, \dots, i\}$ within the first i turns, we define relative preference scores $\mathcal{S}_{\text{real}}^{(j)}$ and $\mathcal{S}_{\text{ref}}^{(j)}$, corresponding to the currently generated response and a previously generated response

retrieved from the augmentation database, respectively. Both are computed via:

$$\mathcal{S}^{(j)} = \mathcal{C}_{\text{avg}[2]}^{(j)} - \mathcal{C}_{\text{avg}[0]}^{(j)} \quad (13)$$

where $\mathcal{C}_{\text{avg}[2]}^{(j)}$ and $\mathcal{C}_{\text{avg}[0]}^{(j)}$ represent positive and negative sentiment, respectively.

To aggregate over a multi-turn trajectory, we apply a discount factor γ :

$$\mathcal{S}_{\text{t-real}}^{(i)} = \sum_{j=1}^i \gamma^{j-1} \cdot \mathcal{S}_{\text{real}}^{(j)}, \quad \mathcal{S}_{\text{t-ref}}^{(i)} = \sum_{j=1}^i \gamma^{j-1} \cdot \mathcal{S}_{\text{ref}}^{(j)} \quad (14)$$

The sentiment gap $\mathcal{S}_{\text{t-ref}}^{(i)} - \mathcal{S}_{\text{t-real}}^{(i)}$ then acts as the advantage signal, allowing TraceBias to capture evolving sentiment and weight dialogue turns unequally.

3.3.3 Final Representation of TraceBias

TraceBias combines trajectory-level signals with DAM-normalized policy dynamics to define its training objective. Integrating the trajectory-based advantages with normalized ratios yields:

$$\mathcal{L}_{\text{final}} = - \sum_{i=1}^N \left(\mathcal{S}_{\text{t-ref}}^{(i)} - \mathcal{S}_{\text{t-real}}^{(i)} \right) \cdot \pi_{\text{ratio}}^{(i)} \quad (15)$$

This objective enables stable, fine-grained updates without relying on binary preference labels, forming a robust optimization basis for the broader ARF framework. (For further mathematical analysis, see Appendix B.3.)

4 Experiments

4.1 Experimental Setup

We systematically evaluate our proposed **ARF-RLHF** framework against widely used baselines, including **DPO** (Rafailov et al., 2024), **PPO** (Schulman et al., 2017b), **RLAIF** (Lee et al., 2024), and **KTO** (Ethayarajh et al., 2024). All models are fine-tuned using **LoRA** (Hu et al., 2022), which provides parameter-efficient adaptation without compromising performance.

Our evaluation spans four representative lightweight LLMs—**Gemma2-2B** (Team et al., 2024), **Qwen2-1.5B** (Yang et al., 2024), **Qwen2.5-1.5B** (Qwen et al., 2025), and **LLaMA3.2-3B** (Grattafiori et al., 2024)—ensuring that findings are not specific to a single architecture or training recipe. We further cover five diverse datasets

from Big-Bench (Srivastava et al., 2023), targeting different capabilities: **Alpaca** (Taori et al., 2023) (instruction following), **GSM8K** (Cobbe et al., 2021) (mathematical reasoning), **StrategyQA** (Geva et al., 2021) (commonsense reasoning), **TopicalChat** (Gopalakrishnan et al., 2023) (dialogue), and **CNN/DailyMail** (See et al., 2017) (summarization). This combination allows us to test both reasoning-heavy and open-domain conversational scenarios. In addition, we include evaluation on the **HH-RLHF** human-annotated dataset (Bai et al., 2022), which serves as a high-quality benchmark for alignment under real human feedback and enables direct comparison with prior studies such as KTO.

Additionally, we introduce a large-scale preference dataset, **Emotion3**, which integrates 78,630 samples from DailyDialog (Li et al., 2017), GoEmotions (Demszky et al., 2020), ISEAR (Scherer and Wallbott, 1997), and Sentiment140 (Go et al., 2009). Compared with typical RLHF datasets that are small and domain-limited, Emotion3 is deliberately broad and noisy, reflecting the variability of real world user feedback. The experiments related to satisfaction are conducted on Emotion3 or its subsets.

All details of dataset construction, validation, and experimental hyperparameter settings are provided in Appendix H, Appendix E and Appendix A.

This experimental design enables us to probe the following three questions:

- Q1** Can ARF reliably collect and adapt to dynamic user preferences, especially under noisy or shifting conditions?
- Q2** How does TraceBias compare with PPO, DPO, RLAIIF and KTO when evaluated under consistent and fair reward supervision?
- Q3** Which specific mechanisms in ARF-RLHF (e.g., ER, DAM, rescaling) are essential for stability and generalization?

4.2 Q1: Can ARF Reliably Track Preferences?

A key challenge in RLHF is whether feedback signals can be both **accurate** and **adaptable** over time. Static scorers provide stable supervision but may fail once user preferences shift, whereas dynamic scoring risks instability. We therefore first test ARF’s reliability in both static and dynamic scenarios.

Dataset	DailyDialog	GoEmotions	ISEAR	Sentiment140	Emotion3
Accuracy (%)	70.05	73.65	76.00	74.10	71.60

Table 1: Test accuracy of the static satisfaction scorer. Reliable static signals are a prerequisite for effective RLHF.

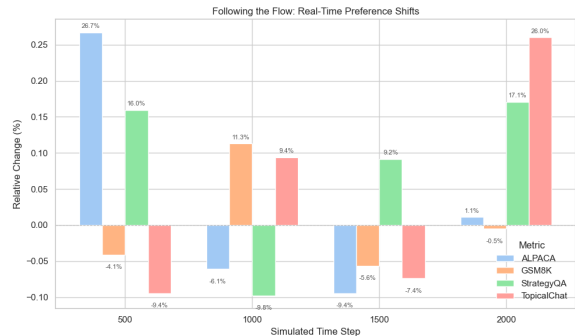


Figure 3: Tracking preference shifts using ARF over simulated time steps. Performance drops reflect deliberate adaptation to new negative signals, validating robustness under non-stationary feedback.

Static supervision quality. Table 1 reports the accuracy of our static satisfaction scorer across five sentiment benchmarks derived from Emotion3. The consistently high accuracy (all >70%) demonstrates that ARF can rely on this module as a trustworthy backbone. This step is crucial: without verifying the base quality of preference signals, subsequent alignment results would be difficult to interpret.

Adaptive preference tracking. We then examine whether ARF can dynamically adapt to shifting preferences by sequentially injecting bias-specific data every 500 steps (ALPACA → GSM8K → StrategyQA → TopicalChat), leveraging the static scorer’s accuracy as a reference. When more than two biases overlap, negative supervision is applied to the earliest bias. As shown in Figure 3, ARF successfully tracks these shifts, with clear downward adjustments once negative preferences are introduced. Initially, synchronized gain/loss patterns emerge between ALPACA and StrategyQA, and between GSM8K and TopicalChat, likely due to semantic similarity; nevertheless, ARF rapidly disentangles these correlations once negative supervision is applied (e.g., ALPACA at step 1500), avoiding stale alignment and demonstrating robustness to preference contradictions—a scenario often overlooked in prior RLHF research.

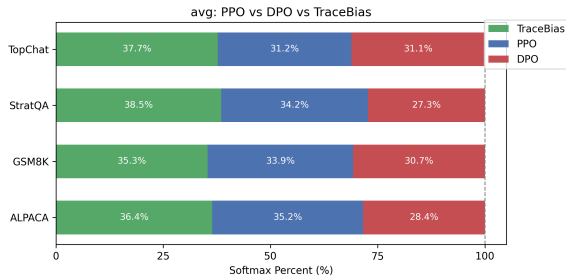


Figure 4: Average performance comparison under different baselines’ fine-tuning. TraceBias consistently outperforms PPO and DPO across tasks. Single models’ performance in appendix J.

4.3 Q2: How Does TraceBias Compare Against RLHF Baselines?

Most prior work evaluates PPO and DPO using multiple LLM judges, but these methods are sensitive to prompt phrasing and evaluation bias. To ensure fairness, we employ a **unified reward model** both for preference filtering and performance evaluation (See more in discussion 5.1). This allows us to ask: under identical supervision, which method learns most effectively?

Performance across tasks. Figure 4 shows that TraceBias consistently outperforms both PPO and DPO across all tasks and model sizes, with average improvements of +3.3% over PPO and +7.6% over DPO. These gains are not marginal: in low-resource RLHF settings, even small percentage improvements can determine whether the aligned model is useful in practice. We attribute the improvements primarily to **Double Average Method (DAM)**, which stabilizes training updates and prevents reward hacking.

Robustness under LLM-based preferences. Human preference data are expensive and often limited, making synthetic labels an appealing alternative. We construct a 1K StrategyQA preference dataset using DeepSeek-v3 (DeepSeek-AI et al., 2025b). As shown in Table 2, TraceBias outperforms both RLAIIF-PPO and RLAIIF-DPO under this synthetic supervision. Interestingly, DPO performs slightly better than PPO here, likely because the preference signals are clearer—suggesting it benefits from more finely curated datasets. Together with the results in the *performance across tasks* section, these findings show that TraceBias delivers superior robustness, consistently maintaining strong performance across diverse bias conditions and supervision regimes.

Evaluation method	RLAIIF-PPO	RLAIIF-DPO	TraceBias
Preference Score (%)	30.3	32.8	36.9
DeepSeek-V3 (win rate vs. Base)	43%	49%	52%

Table 2: Comparison of RLAIIF variants and TraceBias on StrategyQA with DeepSeek-v3 preferences. TraceBias achieves the highest win rate, confirming robustness under noisy labels.

Evaluation method	DPO	KTO	TraceBias
Preference Score (%)	34.2	31.7	34.1
GPT-5 (win rate vs. Base)	57%	51%	54%

Table 3: Comparison of DPO, KTO, and TraceBias on a real human-annotated dataset HH-RLHF.

Evaluation on human-annotated data. To further verify robustness in realistic scenarios, we employ a hybrid validation strategy that combines human annotations with LLM-based judgments from GPT-5 (OpenAI, 2025), confirming that the unified scorer remains a reliable basis for evaluation. Based on this setup, we additionally evaluate DPO, KTO, and TraceBias on a real human-labeled dataset. As shown in Table 3, TraceBias again achieves competitive performance, slightly outperforming KTO and closely matching DPO. Notably, DPO shows stronger performance under high-quality human annotations, which is consistent with prior findings in the KTO paper (Section 4.5: *KTO vs. DPO – when to use which?*). This observation aligns with our earlier discovery: while KTO is more robust under noisy or synthetic preference labels, DPO tends to excel when reliable human feedback is available, and TraceBias balances both regimes effectively. Importantly, even when all three methods are evaluated on the same human-annotated dataset, their relative advantages remain consistent.

4.4 Q3: Which Design Components Are Critical?

While ARF-RLHF introduces several innovations, it is important to understand which components are indispensable. We therefore ablate **Experience Replay (ER)**, **Double Average Method (DAM)**, and **rescoring**.

Experience Replay (ER). Table 4 shows that removing ER (ratio = 0) yields higher short-term accuracy on recent tasks but significantly degrades performance on the Emotion3 dataset, indicating poor generalization. Conversely, a moderate ER ratio (0.5) balances adaptation and long-term ro-

ER Ratio	GSM8K (accuracy)	Emotion3 (accuracy)
Basic	53.52	73.84
0	60.59	59.32
0.5	56.40	70.88

Table 4: Ablation of ER ratio. A moderate setting improves generalization and yields higher accuracy, whereas no replay causes catastrophic forgetting with a sharp drop in accuracy.

business. This suggests that ER acts as a memory mechanism, mitigating catastrophic forgetting when preferences evolve—a property often overlooked in existing RLHF pipelines.

Gradient stability. Figure 2 compares gradient dynamics on Qwen2.5-1.5B fine-tuned over GSM8K. While PPO suffers from high variance even with clipping, TraceBias with DAM yields both lower magnitude and variance. Stable gradients are crucial in RLHF, where unstable updates can quickly derail alignment. This validates DAM as a key factor behind TraceBias’s performance.

Rescoring necessity. As shown earlier in Table 5, the ablation study compares the use and omission of ReScoring within experiment in *Adaptive Preference Tracking* via ARF. When ALPACA and GSM8K undergo preference reversal, removing rescoring results in inflated scores and persistent misalignment. This demonstrates that rescoring is not merely an auxiliary option but an essential mechanism to ensure preference signals remain valid over time.

4.5 Summary of Findings

In summary, our experiments show that ARF reliably captures preference signals and adapts to evolving supervision even under noise or conflicts, TraceBias matches the performance of PPO and DPO across both human-annotated and synthetic preferences, and core components such as ER, DAM, and rescoring are indispensable for stable training and long-term alignment. Detailed case studies can be found in Appendix K.

5 Discussion

5.1 How can we prove the accuracy of our experiments?

As noted in Section 4.3, AI-agent-based evaluation (e.g., using an LLM judge) exhibits high variance from prompt wording, task quirks, model ar-

Condition	ALPACA ($\Delta\%$)	GSM8K ($\Delta\%$)
With ReScore	-9.4	-0.5
Without ReScore	7.8	3.7

Table 5: Relative change ($\Delta\%$) on ALPACA and GSM8K after preference reversal. Negative values indicate successful adaptation.

chitecture, and random seeds, leading to inconsistent results. We therefore focus on comparing the convergence of different RLHF methods under identical preference supply. To mitigate variance and ensure real-world applicability, we combine AI-agent metrics (Section 4.3) with a unified scorer: reward-oriented methods (e.g., TraceBias, PPO) train against a fixed reward model, while comparison-based methods (e.g., DPO, KTO) use the same scorer for alignment. All runs share the same held-out test set (unseen in training) and immutable scorer, preventing coupling or leakage. In RLAIIF and HH-RLHF experiments, we further employ two distinct AI-agent evaluations (DeepSeek-V3 and GPT-5); though less reproducible, they provide complementary evidence that our scorer remains valid in realistic settings, enabling stable, unbiased comparison of convergence to target preferences.

5.2 On the Reliability of ARF Satisfaction Supervision

Satisfaction annotations are inherently subjective-labeling tasks involving large numbers of human annotators often reflect diverse preferences, even SOTA LLMs show bias when re-labeling Emotion3, with neutral predictions ranging from 24.0% to 37.3% (variance 29.51, Appendix G). Neutrality proves especially ambiguous, leading to unstable supervision. Although our static scorer reaches only 70% accuracy, it mitigates such uncertainty by excluding neutral scores during TraceBias updates (Eq. 13), using them instead to dampen noisy gradients e.g., Good: 0.02, Neutral: 0.90, Bad: 0.08 yields $S = -0.06$ resulting in minimal updates. This design—by ensuring softer updates for items dominated by neutral evaluations—avoids rigid binary comparisons under ambiguity, thereby improving robustness, reducing annotation variance, and enabling reliable tracking of individual preferences beyond crowd-level bias.

6 Conclusion

We introduce **ARF-RLHF**, a reinforcement learning framework that aligns language models to individual user preferences. By modeling continuous satisfaction and adapting to evolving feedback, it enables personalized, context-aware behavior beyond population-level signals. Experiments show robust preference optimization under limited supervision, offering a scalable path to adaptive models that reflect individual intent while maintaining rigor and reliability.

7 Potential Risks

While ARF-RLHF introduces a scalable and adaptive framework for preference alignment, its deployment in real world applications may raise several practical considerations.

1. **System Integration and Stability.** The framework introduces multiple interacting components, including feedback processing, scoring modules, and periodic model updates. In production environments, coordinating these components may increase system complexity and require careful engineering to ensure stable and reliable operation.
2. **Increased Computational Overhead.** Compared to standard inference-only pipelines, ARF-RLHF introduces additional components such as feedback processing, scoring, and periodic model updates. These steps may lead to higher computational costs and latency in production environments, particularly under large-scale or real-time usage scenarios.

Overall, these considerations do not constitute fundamental limitations, and can be addressed through thoughtful system design and deployment practices, enabling practical and responsible adoption of ARF-RLHF in real world settings.

8 Limitations

While our method offers significant theoretical and empirical advantages, we acknowledge the following limitations:

1. **Lack of Real Human evaluation:** The scorer-based evaluation protocol is designed for fair comparison of RLHF methods, focusing on minimizing noise factors. It does not yet capture generalization performance in real world scenarios. Future work will include human blind testing or cross-validation with

alternative scorers to further validate the results.

2. **Model Scale Constraint:** Due to resource constraints, we conduct experiments on 1.5B–3B models. While our method is architecture-agnostic and theoretically scalable, its effectiveness on larger LLMs (e.g., 7B, 13B, 65B) remains to be explored in future work.

9 Use of AI Assistants

AI assistants (e.g., ChatGPT) were used to support phrasing, grammar, and clarity improvements during manuscript preparation. All research ideas, methods, experiments, and conclusions were developed by the authors. All AI assisted content was carefully reviewed and verified by the authors to ensure correctness and originality.

References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- RALPH ALLAN BRADLEY. 1955. [Rank analysis of incomplete block designs: Iii. some large-sample results on estimation and power for a method of paired comparisons*](#). *Biometrika*, 42(3-4):450–470.
- Huanpei Chen and Hsin-Hsi Chen. 2016. [Implicit polarity and implicit aspect recognition in opinion mining](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.

- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [Goemotions: A dataset of fine-grained emotions](#). *Preprint*, arXiv:2005.00547.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. [Kto: Model alignment as prospect theoretic optimization](#). *Preprint*, arXiv:2402.01306.
- Beverly Fehr and James A. Russell. 1984. [Concept of emotion viewed from a prototype perspective](#). *Journal of Experimental Psychology: General*, 113(3):464–486.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies](#). *Preprint*, arXiv:2101.02235.
- Alec Go, Richa Bhayani, and Lei Huang. 2009. [Twitter sentiment classification using distant supervision](#). Technical Report CS224N Project Report, Stanford University.
- Karthik Gopalakrishnan, Behnam Hedayatnia, Qianlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. [Topical-chat: Towards knowledge-grounded open-domain conversations](#). *Preprint*, arXiv:2308.11995.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Alastair Henry, Cecilia Thorsen, and Peter D. MacIntyre. 2021. [Willingness to communicate in a multilingual context: part two, person-context dynamics](#). *Journal of Multilingual and Multicultural Development*, 42(9):827–841.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. [Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback](#). *Preprint*, arXiv:2309.00267.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- OpenAI. 2025. [Gpt-5 system card](#). <https://openai.com/research/gpt-5-system-card>. Accessed: 2025-09-20.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Shrimai Prabhumoye, Sushant Choudhary, Eleni Spiliopoulou, Chris Bogart, Carolyn Rose, and Alan W Black. 2017. [Linguistic markers of influence in informal interactions](#). In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 53–62, Vancouver, Canada. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. [Direct preference optimization: Your language model is secretly a reward model](#). *Preprint*, arXiv:2305.18290.
- Klaus R. Scherer and Harald G. Wallbott. 1997. ISEAR dataset: International survey on emotion antecedents and reactions. <https://www.unige.ch/cisa/research/materials-and-online-research/research-material/>.
- John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2017a. [Trust region policy optimization](#). *Preprint*, arXiv:1502.05477.
- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2018. [High-dimensional](#)

continuous control using generalized advantage estimation. *Preprint*, arXiv:1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017b. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

James Shanahan, Yan Qu, and Janyce Wiebe. 2006. *Computing Attitude and Affect in Text: Theory and Applications*, volume 20.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*. Featured Certification.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Yla R. Tausczik and James W. Pennebaker. 2010. [The psychological meaning of words: Liwc and computerized text analysis methods](#). *Journal of Language and Social Psychology*, 29(1):24–54.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.

Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Haorui Wang, Zhen Qin, Feng Han, Jialu Liu, Simon Baumgartner, Michael Bendersky, and Chao Zhang. 2024. [PLaD: Preference-based large language model distillation with pseudo-preference pairs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15623–15636, Bangkok, Thailand. Association for Computational Linguistics.

A Hyperparameters

All experiments’ Hyperparameters shows below:

1. Static Scorer Evaluation: table 9
2. Adaptive Preference Tracking: table 10
3. RLHF Comparison: table 11
4. RLAIIF vs. ARF: table 12
5. Effect of Experience Replay (ER) in ARF 13
6. Gradient Stability Analysis for DAM 14

All experiments were conducted on 2 * NVIDIA GTX 2080 Ti GPUs with CUDA unified memory enabled, using multi-GPU parallelism. Training consumed approximately 390 GPU-hours per device.

B The Essential Homology of DPO, PPO, and TraceBias

The Actor-Critic (AC) algorithm can be formulated as:

$$\mathcal{L}^{\text{AC}}(\theta) = - \sum_{t=1}^N \log \pi_{\theta}(a_t | s_t) \cdot A_t \quad (16)$$

In this section, we demonstrate the theoretical connections among PPO, DPO, and our proposed TraceBias. We argue that these methods share a common optimization structure, thereby validating the theoretical soundness of TraceBias.

B.1 PPO as a Clip-Constrained Actor-Critic Variant

PPO (Schulman et al., 2017b) can be defined as:

$$\mathcal{L}^{\text{PPO}}(\theta) = \mathbb{E}_t [\min (r_t(\theta)A_t, \text{clip}(r_t(\theta), \epsilon)A_t)] \quad (17)$$

where $r_t(\theta)$ is the importance sampling ratio between the current and previous policies:

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \quad (18)$$

and the clipping function is defined as:

$$\text{clip}(r_t(\theta), \epsilon) = \begin{cases} 1 - \epsilon, & \text{if } r_t(\theta) < 1 - \epsilon \\ r_t(\theta), & \text{if } 1 - \epsilon \leq r_t(\theta) \leq 1 + \epsilon \\ 1 + \epsilon, & \text{if } r_t(\theta) > 1 + \epsilon \end{cases} \quad (19)$$

By expanding the objective, we obtain:

$$\mathcal{L}^{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min \left(\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} A_t, \text{clip} \left(\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, \epsilon \right) A_t \right) \right] \quad (20)$$

If we ignore the clipping operation—which serves as a regularization mechanism to limit the policy update magnitude—the PPO objective reduces to:

$$\mathcal{L}^{\text{PPO}}(\theta) \propto \sum_{t=1}^N r_t(\theta) A_t \quad (21)$$

Here, $r_t(\theta)$ reflects the policy ratio $\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}$, which encourages increasing the likelihood of actions with high advantage A_t . This shows that PPO essentially shares the same optimization direction as AC, differing only in the incorporation of a trust-region-inspired constraint to stabilize training.

B.2 DPO as a KL-Constrained Actor-Critic Variant

Direct Preference Optimization (DPO) (Rafailov et al., 2024) leverages the Bradley-Terry model to represent pairwise preferences as follows:

$$\mathbb{P}(y^+ \succ y^- | x) = \frac{\exp(r(y^+))}{\exp(r(y^+)) + \exp(r(y^-))} \quad (22)$$

Its associated loss is:

$$\mathcal{L}_R(\phi, D) = -\mathbb{E}_{(x, y_w, y_l)} \left[\log \frac{\exp(r(y_w))}{\exp(r(y_w)) + \exp(r(y_l))} \right] \quad (23)$$

The DPO objective derived from this model is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x) \pi_{\text{ref}}(y_l | x)}{\pi_\theta(y_l | x) \pi_{\text{ref}}(y_w | x)} \right) \right] \quad (24)$$

Introducing a normalization constant $Z(x)$, the implicit optimal policy $\pi^*(y | x)$ can be defined as:

$$\pi^*(y | x) = \frac{\pi_{\text{ref}}(y | x) \cdot \exp(\frac{1}{\beta} r(y))}{Z(x)} \quad (25)$$

where the partition function $Z(x)$ is:

$$Z(x) = \sum_{y'} \pi_{\text{ref}}(y' | x) \cdot \exp \left(\frac{1}{\beta} r(y') \right) \quad (26)$$

Taking the logarithm of both sides yields:

$$\log \pi^*(y | x) = \log \pi_{\text{ref}}(y | x) + \frac{1}{\beta} r(x, y) - \log Z(x) \quad (27)$$

We can then derive:

$$r(x, y) = \beta \log \frac{\pi^*(y | x)}{\pi_{\text{ref}}(y | x)} + \beta \log Z(x) \quad (28)$$

By applying $r(x, y)$ in the pairwise preference model $\mathbb{P}(y^+ \succ y^- | x)$ and utilizing the Plackett-Luce model (see Appendix A.3 of (Rafailov et al., 2024) for more details), the DPO objective can be equivalently rewritten as:

$$\max_{\pi_\theta} \left\{ \mathbb{E}_{x \sim D, y \sim \pi_\theta(\cdot | x)} [r_\phi(x, y)] - \beta \text{KL}(\pi_\theta(\cdot | x) \| \pi_{\text{ref}}(\cdot | x)) \right\} \quad (29)$$

Here, the optimization objective is to maximize the expected reward regularized by a KL divergence term. Assuming $A(x, y) = r(x, y)$, and temporarily ignoring the KL regularization, this reduces to an actor-critic style objective:

$$\mathcal{L}^{\text{DPO}}(\theta) \propto \mathbb{E}_{y \sim \pi_\theta(\cdot | x)} [A(x, y)] \quad (30)$$

This implies that DPO and actor-critic share effectively the same optimization objective when the reward signal is defined as the advantage. In practice, this expectation can be approximated by:

$$\mathcal{L}^{\text{DPO}}(\theta) \approx \sum_{i=1}^N \log \pi_\theta(y_i | x_i) \cdot A(x_i, y_i) \quad (31)$$

Thus, DPO can be interpreted as a KL-regularized actor-critic method, where the reward signal is derived from preference feedback rather than scalar returns.

B.3 TraceBias as a DAM-Constrained Actor-Critic Variant

We have previously outlined the Actor-Critic (AC) interpretation of TraceBias in our methodology. Here, we provide a formal derivation from the expanded formulation to its AC-style representation.

$$\mathcal{L}_{\text{final}} = - \sum_{i=1}^N (\mathcal{S}_{\text{t-ref}} - \mathcal{S}_{\text{t-real}}) \cdot \exp(\pi_\theta^{(i)} - \pi_{\text{old}}^{(i)}) \quad (32)$$

Following our methodology, the token-level reward (or *score*) is defined via discounted(γ) step-wise preferences as follows:

$$\mathcal{S}_{\text{t-real}}^{(i)} = \sum_{j=1}^i \gamma^{j-1} \cdot \mathcal{S}_{\text{real}}^{(j)}, \quad \mathcal{S}_{\text{t-ref}}^{(i)} = \sum_{j=1}^i \gamma^{j-1} \cdot \mathcal{S}_{\text{ref}}^{(j)} \quad (33)$$

Accordingly, inspired by Generalized Advantage Estimation (GAE) (Schulman et al., 2018), we define the advantage function by measuring the difference between the real and reference trajectories:

$$\mathcal{A}_i = \mathcal{S}_{\text{t-ref}}^{(i)} - \mathcal{S}_{\text{t-real}}^{(i)} \quad (34)$$

Substituting this into the objective, TraceBias can be rewritten in an actor-critic form:

$$\mathcal{L}_{\text{TraceBias}} = - \sum_{i=1}^N \mathcal{A}_i \cdot \exp(\pi_{\theta}^{(i)} - \pi_{\text{old}}^{(i)}) \quad (35)$$

To improve optimization stability, we introduce the DAM smooth surrogate strategy, which pools token-level scores and normalizes the policy ratio across the trajectory. Analogous to the clipping term in PPO and the KL regularization in DPO, DAM serves as a regularization mechanism: We define the token-level ratio as:

$$\pi_{\theta}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_{\theta}(y_t^{(i)} | x^{(i)}) \quad (36)$$

$$\pi_{\text{old}}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_{\text{old}}(y_t^{(i)} | x^{(i)}) \quad (37)$$

Then, building upon the improvements from TRPO (Schulman et al., 2017a) and PPO (Schulman et al., 2017b), the normalized trajectory-level policy ratio is computed as:

$$\pi_{\text{ratio}}^{(i)} = \exp(\pi_{\theta}^{(i)} - \pi_{\text{old}}^{(i)}) \quad (38)$$

Thus, the final form of TraceBias is:

$$\mathcal{L}_{\text{TraceBias}} = - \sum_{i=1}^N \mathcal{A}_i \cdot \pi_{\text{ratio}}^{(i)} \quad (39)$$

This derivation shows that TraceBias can be directly interpreted as an Actor-Critic method without introducing additional approximations, highlighting its theoretically grounded and streamlined formulation.

Scale value	Scale Impact on Sigmoid under 0.2	interval
1(Too small)	0.5498	[0.3775,0.6225]
4	0.6900	[0.1192,0.8808]
6	0.7685	[0.0474,0.9526]
8	0.8320	[0.018,0.9820]
16(Too big)	0.9608	[0.0003,0.9997]

Table 6: The table shows how different scale values map to specific intervals after applying the Sigmoid function.

B.4 Summary

Although PPO, DPO, and TraceBias differ significantly in their final objectives, all can be reformulated as variants of the Actor-Critic (AC) framework. By optimizing the expected reward weighted by advantage, each method introduces distinct regularization strategies—such as PPO’s clipping, DPO’s KL constraint, or TraceBias’s DAM normalization—to improve training stability. This unified perspective highlights that TraceBias is not only theoretically grounded but also competitive with existing policy optimization techniques.

C The pseudocode of TraceBias

The TraceBias pseudocode shows in alg 1.

D The Selection of Sigmoid Scale

Scale plays a crucial role in determining the sensitivity of the norm function, as it directly affects most parameters. We recommend selecting scales within the range of [4,8], as excessively large values can push most parameters towards the boundaries of the Sigmoid function interval, limiting their effective range. Table 6 illustrates when $(\text{CosSim}(\mathcal{C}_{\text{ARF_avg}}, \mathcal{C}_{\text{basic_avg}}) - 0.5)$ equals values 0.2 is converted together with the corresponding function value range by the Sigmoid function under different proportional settings. It is worth noting that when the scale is set to 1, the relatively large 0.2 level in sigmoid results in a mapping value of only 0.5498. However, when the ratio increases to 16, the same input is mapped to nearly 1, indicating that the range is overly compressed. Based on these observations, we strongly recommend selecting a scale within the range of [4,8] for the numerical deviation annotation algorithm, as it ensures a balanced transformation without pushing values to extremes.

E RLHF Dataset Construction

To support comparison-based fine-tuning methods such as DPO and PPO, we construct a simulated

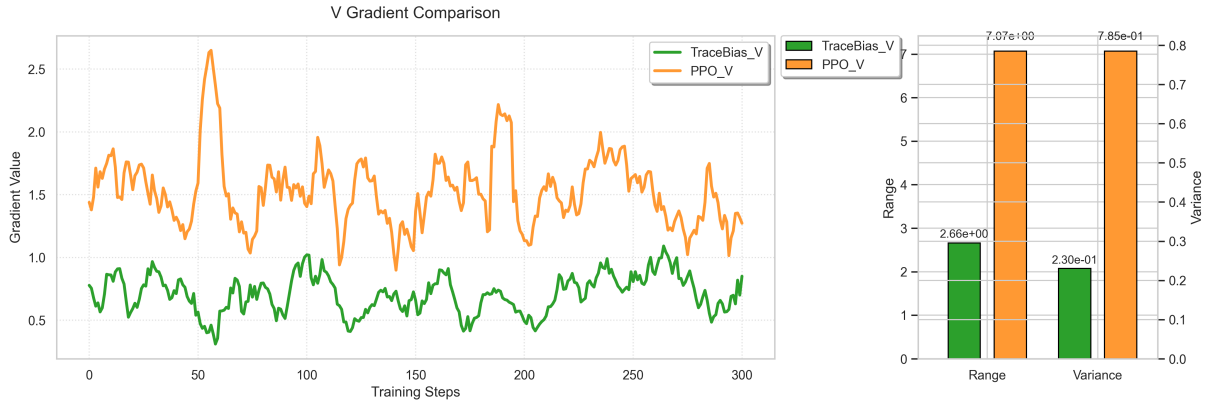


Figure 5: V Gradient norm comparison between PPO (with clip range $\epsilon = 0.2$) and TraceBias with DAM.

binary preference dataset. Given the prohibitive cost of large-scale human annotation, and the fact that this dataset is primarily used to compare fine-tuning preferences across methods rather than for real-world deployment, we adopt a surrogate construction strategy that also aligns with the training of our ARF preference model.

Concretely, we employ the `nlpaug.augmenter.word.SynonymAug` module from the `nlpaug` library to perform four rounds of synonym substitution using WordNet, generating paraphrased variants that preserve semantic intent while introducing surface-level diversity.

Before constructing the binary comparison dataset, to amplify preference signals and avoid potential overfitting, we prepend task-specific prompts that were never included in the scorer’s training data. We annotate the augmented **’former’** samples with soft labels using our fine-tuned static satisfaction scorer. Noise instances are sourced from the CNN/DailyMail corpus. Specifically:

- **Good prompt:** Great! You gave a correct answer. Here is the next question: ...
- **Bad prompt:** Your answer is absolutely wrong! This is the next question. Stop giving such terrible and misleading feedback! ...

The ARF scorer is then fine-tuned on pairs of the preceding dialogue (inputs) and the satisfaction scores (labels) annotated from the simulated responses, serving as a proxy for user-aligned preferences. To ensure fairness, each RLHF method (including PPO and DPO) constructs its training pairs using the same ARF scorer: for each pair, the sample with a higher score is designated as the preferred (positive) response, while the lower-scored

Evaluation method	Win	Lose	Equal
RLAIF-PPO	43%	50%	7%
RLAIF-DPO	49%	47%	4%
TraceBias	52%	44%	4%

Table 7: The win, lose, equal rate compare to basic model

one is treated as negative. This guarantees that all methods are aligned in their optimization direction and evaluated under consistent supervision.

F Evaluation of DeepSeek Agent under the RLAIF Task

We use below prompt to compare the output of baselines and basic models, the win/lose/equal table shows in table 7. The prompt of comparison shows below:

Question: ‘...’ Answer1: ‘...’ Answer2: ‘...’ Please use strict criteria to determine which answer is more in line with human preferences 1 or 2 if it is similar give 3 only answer a number.

G Table of Subjective Experiments

The table 8 reveals that even state-of-the-art LLMs exhibit significant variance in emotion judgment, especially in distinguishing neutral sentiment—underscoring the inherent noise in satisfaction estimation.

H Satisfaction Dataset Construction

To construct a large-scale, diverse, and high-quality satisfaction classification dataset aligned with our three-level labeling schema (*bad*, *neutral*, *good*),

Model	Positive (%)	Neutral (%)	Negative (%)
Actual Samples (Uniform)	33.3	33.3	33.3
GPT-4o	40.3	30.3	29.3
DeepSeek-v3	44.0	24.0	32.0
Gemini 2.0 Flash	35.0	37.3	27.6
Variance	13.64	29.51	13.28

Table 8: Variance of label distribution across the three models on Emotion3 samples, calculated relative to the mean across all three models. Neutral sentiment shows the greatest spread, highlighting inconsistency in satisfaction estimation.

we aggregate a total of 78,630 samples from four widely-used emotion and sentiment datasets:

- **DailyDialog** (Li et al., 2017): A multi-turn dialogue dataset that closely mirrors everyday conversational scenarios.
- **GoEmotions** (Demszky et al., 2020): A fine-grained, high-quality emotion classification dataset spanning a wide range of affective states.
- **ISEAR** (Scherer and Wallbott, 1997): A clean and structured emotion dataset based on psychological self-reports.
- **Sentiment140** (Go et al., 2009): A large-scale Twitter sentiment dataset that reflects informal and noisy online communication.

To unify the labeling across datasets with heterogeneous annotation schemes, we define a common strength-based mapping strategy, converting existing emotion tags into a standardized 7-level satisfaction scale (see Table 15). For relatively clean datasets (DailyDialog, GoEmotions, ISEAR), we directly apply this mapping to assign satisfaction scores.

Given the informal nature of Sentiment140, additional cleaning is necessary. We sample 15,000 instances and perform multi-round evaluation using both Qwen2 7B and LLaMA3 13B. Each sample is scored twice by each model; the maximum and minimum scores are discarded, and the mean of the remaining two is taken as the final label. Samples with high variance across scores are further manually verified to ensure annotation reliability. The result is a cleaned subset of 15,000 samples from Sentiment140 with stable satisfaction labels.

After consolidating all datasets, we create a unified **Emotion7** dataset with 7 satisfaction levels. We then perform a coarse mapping to form the final **Emotion3** dataset: levels [0, 1] as *bad*, 1.5 as *neutral*, and [2, 3] as *good*. This dataset provides broad domain coverage, consistent labels, and stylistic diversity, serving as the basis for training our static

satisfaction scorer.

I Gradient Norm Comparison

Figure 2 and Figure 5 show the gradient norm comparison between PPO (clip) and TraceBias (DAM). Both Q and V curves are smoothed using a moving-average window of size 10.

Here, Q and V exhibit similar temporal patterns during training, while differing in dynamic range and noise variance. Given that Q, K, and V are derived from the same input representations and jointly influence the attention output, updates to these components are naturally coupled. Since the gradient norm abstracts away directional information and captures only magnitude, such coupling can manifest as similar temporal trends.

Overall, consistently lower gradient norms and reduced variance suggest improved optimization stability.

J Models' Performance under different RLHF Baselines

We show all models' RLHF performance below:

- Qwen2-1.5B: figure 6
- Qwen2.5-1.5B: figure 7
- Gemma2-2B: figure 8
- Llama3.2-3B: figure 9

We applied softmax with temperature (set to 0.1) purely for visualization purposes.

K Case Study under Llama3.2

To preserve the original formatting of model outputs, we retain their format in the paper. For excessively long responses, we replace parts with ellipses ("...") for clarity. Representative examples are provided in Table 16, Table 17, Table 18, and Table 19.

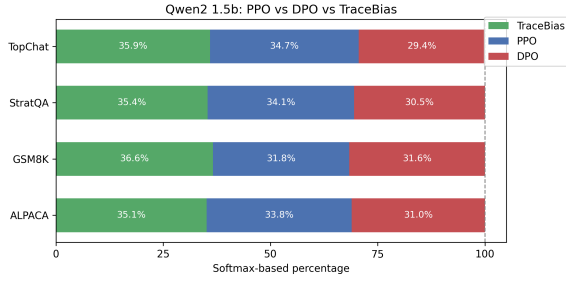


Figure 6: Qwen2’s Performance

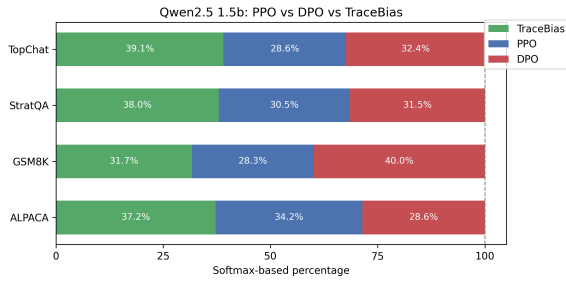


Figure 7: Qwen2.5’s Performance

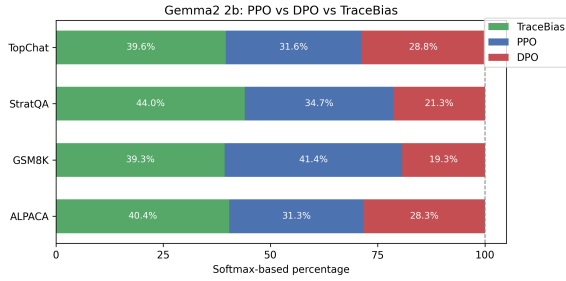


Figure 8: Gemma2’s Performance

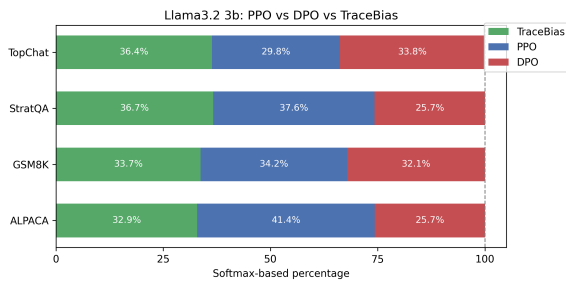


Figure 9: Llama3.2’s Performance

Algorithm 1: TraceBias with DAM

Input: Augmented dialogue dataset \mathcal{D}_{aug} , policy model P_θ , reference model P_{old} , ARF-scorer \mathcal{F}_{ARF} , max length L_{max} , discount factor γ , learning rate η

Output: Updated parameters θ

while training do

repeat

Sample a dialogue trajectory

$\{(x_i, y_i^*, s_{\text{ref}})\}_{i=1}^N \sim \mathcal{D}_{\text{aug}}$;

Initialize total loss $\mathcal{L} \leftarrow 0$;

for $i = 1$ to N **do**

// Generate model output for turn i

$y_i \sim P_\theta(\cdot | x_i)$;

if $\text{length}(y_i) > L_{\text{max}}$ **then**

└ **break**

// DAM-normalized token-level log-probabilities

$\pi_\theta^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_\theta(y_t^{(i)} | x_i)$;

$\pi_{\text{old}}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_{\text{old}}(y_t^{(i)} | x_i)$;

// Initialize accumulated scores for current i

$\mathcal{S}_{\text{t-ref}}^{(i)} \leftarrow 0$;

$\mathcal{S}_{\text{t-real}}^{(i)} \leftarrow 0$;

// Accumulate over each j

within the current i turns

for $j = 1$ to i **do**

// Obtain sentiment distribution from ARF-scorer

$\mathcal{C}_{\text{avg}}^{(j)} \leftarrow \mathcal{F}_{\text{ARF}}(y_j)$;

// Compute per-turn sentiment-based score (Eq. 13)

$\mathcal{S}_{\text{real}}^{(j)} \leftarrow \mathcal{C}_{\text{avg},2}^{(j)} - \mathcal{C}_{\text{avg},0}^{(j)}$;

// Reference score accumulation

(label-based)

$\mathcal{S}_{\text{t-ref}}^{(i)} += \gamma^{j-1} \cdot \mathcal{S}_{\text{ref}}^{(j)}$;

// Accumulate discounted real trajectory score

$\mathcal{S}_{\text{t-real}}^{(i)} += \gamma^{j-1} \cdot \mathcal{S}_{\text{real}}^{(j)}$;

// Importance sampling ratio (Eq. 12)

$\pi_{\text{ratio}}^{(i)} = \exp(\pi_\theta^{(i)} - \pi_{\text{old}}^{(i)})$;

// Add weighted advantage to loss (Eq. 15)

$A^{(i)} \leftarrow \mathcal{S}_{\text{t-ref}}^{(i)} - \mathcal{S}_{\text{t-real}}^{(i)}$;

$\mathcal{L} += -\pi_{\text{ratio}}^{(i)} \cdot A^{(i)}$;

until valid sample obtained;

// Gradient update

Backpropagate: $\nabla_\theta \mathcal{L}$;

Update: $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}$;

Hyper Parameters	TrainStep	Test Step	Learning Rate	Batch	MLP Hidden Size
Static Scorer	20000	500	1e-6	20	328

Table 9: Hyperparameters of Static Scorer Evaluation

Hyper Parameters	TrainStep	Test Step	Learning Rate	ERRatio	MLP Hidden Size
ARF Scorer	2000	500	1e-6	0.5	328

Table 10: Hyperparameters of Adaptive Preference Tracking via ARF

Hyper Parameters	LoRA Rank	Epoch	Test Step/Epoch	Train Step/Epoch	Learning Rate	TraceBias Gamma	Clip Epsilon	PPO [c1,c2]	DPO beta
PPO	8	4	100	500	1e-6	-	0.2	[0.01,0.01]	-
DPO	8	4	100	500	1e-6	-	-	-	0.1
TraceBias	8	4	100	500	1e-6	0.99	-	-	-

Table 11: Hyperparameters for RLHF comparison (KTO hyper: TRL default).

Hyper Parameters	LoRA Rank	Epoch	Test Step/Epoch	Train Step/Epoch	Learning Rate	TraceBias Gamma	Clip Epsilon	PPO [c1,c2]	DPO beta
PPO	8	2	100	250	1e-5	-	0.2	[0.01,0.01]	-
DPO	8	2	100	250	1e-5	-	-	-	0.1
TraceBias	8	2	100	250	1e-5	0.99	-	-	-

Table 12: Hyperparameters of RLAIIF vs. ARF

ERRatio	TrainStep	Test Step	Learning Rate	MLP Hidden Size
0.5	1000	500	1e-5	328
0	1000	500	1e-5	328

Table 13: Hyperparameters of Experience Replay (ER) in ARF

Hyper Parameters	LoRA Rank	Train Step	Learning Rate	TraceBias Gamma	Clip Epsilon	PPO [c1,c2]
PPO	8	300	1e-6	-	0.2	[0.01,0.01]
TraceBias	8	300	1e-6	0.99	-	-

Table 14: Hyperparameters of Gradient DAM test

Strength	Label	DailyDialog	ISEAR	GoEmotion
0.0	extremely neg	disgust	disgust	disappointment embarrassment grief sadness
0.5	neg	anger, sadness	sadness	anger disgust remorse
1.0	neg neutral	fear	fear	annoyance confusion disapproval nervousness
1.5	neutral	no emotion		caring curiosity neutral
2.0	pos neutral			approval desire realization relief surprise
2.5	pos	surprise	joy	admiration amusement excitement joy love optimism pride
3.0	extreme pos	happiness		admiration gratitude

Table 15: Emotion categories mapped to continuous strength scores across datasets.

Table 16: GSM8K

Field	Content
Prompt	Q: Reggie and his brother are having a basketball shooting contest. They each get to take 10 shots. Layups are worth 1 point, free throws are worth 2 points, and anything further away is worth 3 points. Reggie makes 3 layups, two free throws, and one long shot. His brother only shoots long shots and makes 4 of them. How many points does Reggie lose by? A:
label	Reggie makes 3 points off layups because $3 \times 1 = 3$ Reggie makes 4 points off free-throws because $2 \times 2 = 4$ Reggie makes 3 points off long shots because $1 \times 3 = 3$ Reggie scores 10 points because $3 + 4 + 3 = 10$ His brother scores 12 points because $4 \times 3 = 12$ Reggie loses by 2 because $12 - 10 = 2$
Basic model	To find out how many points Reggie lost by, we calculate both totals explicitly: Reggie's points: layups $3 \times 1 = 3$; free throws $2 \times 2 = 4$; long shot $1 \times 3 = 3$; sum = 10. Brother's points: $4 \times 3 = 12$. Since $10 < 12$, Reggie loses by 2 points.
TraceBias	To determine the points Reggie scored, we need to calculate the points from each type of shot. Reggie made 3 layups ($3 \times 1 = 3$ points), 2 free throws ($2 \times 2 = 4$ points), and 1 long shot ($1 \times 3 = 3$ points). Total: $3 + 4 + 3 = 10$ points. His brother made 4 long shots ($4 \times 3 = 12$ points). Difference: $12 - 10 = 2$ points.
Comment	The TraceBias response employ formal mathematical notation and rigorous step-by-step reasoning to enhance precision and clarity.

Table 17: ALPACA

Field	Content
Prompt	Q: Describe the character of Romeo from Romeo and Juliet. A:
label	"Romeo is a passionate, impulsive, and hopelessly romantic young man, who falls deeply in love with Juliet at first sight. He is a member of House Montague in the play <i>Romeo and Juliet</i> , and is depicted as sensitive, idealistic, and poetic, idolizing love and its power. Although he's brave and skilled in combat, Romeo is also shown as prone to intense emotions, leading him to act impulsively and recklessly. His strong feelings often cloud his judgement, which eventually leads to the tragic end of the play. Despite his flaws, Romeo is a loyal friend, devoted son and nephew, and an unwavering lover, willing to do anything for the woman he loves."
Basic model	<p>Romeo Montague is the protagonist of William Shakespeare's tragic play <i>Romeo and Juliet</i>. He is a young, passionate, and impulsive nobleman from the rival Montague family.</p> <ol style="list-style-type: none"> 1. Romantic and passionate: falls in love at first sight and sacrifices everything for Juliet. 2. Impulsive and impetuous: acts on emotion without forethought. 3. Loyal and devoted: fiercely protects family and Juliet. 4. Youthful and inexperienced: mistakes due to naiveté. 5. Emotionally expressive: conveys feelings through poetry. 6. Self-destructive: reckless decisions contribute to tragedy. <p>His tragic fate underscores the destructive power of hate and the redemptive power of love.</p>
TraceBias	Romeo, the protagonist of Shakespeare's tragic love story, is a complex, passionate, and impulsive nobleman from the House of Montague. Key Characteristics: (1) <i>Impulsive and passionate</i> ; (2) <i>Romantic and idealistic</i> ; (3) <i>Impetuous</i> ; (4) <i>Loyal and devoted</i> . Motivations: love for Juliet, family loyalty, desire for vengeance. Flaws: impulsivity, lack of prudence, emotional instability. His character symbolizes love's power and the human spirit.
Comment	The TraceBias response is thorough with the format more close to label.

Table 18: StrategyQA

Field	Content
Prompt	Q: Did compact discs make computer gaming more popular? A:
label	Yes. Compact discs contained significantly more storage space than the previously popular floppy disc format. Gaming studios were therefore able to significantly improve the graphics, sounds, and features of their games to make them more immersive. The better games led to a massive increase in popularity for computer gaming.
Basic model	<p>Compact discs (CDs) did play a role in making computer gaming more popular, but it was not the sole factor.</p> <ul style="list-style-type: none"> • CD-ROM technology: enabled large-game distribution on a single disc. • Improved graphics and sound: enhanced audio-visual experience (e.g. <i>Doom, Myst</i>). • Wider availability: expanded platform reach. • Industry growth: attracted investment and talent. <p>Other influences included hardware advances, online gaming, and marketing strategies.</p>
TraceBias	Yes. CDs significantly popularized PC gaming in the 1990s. The CD-ROM format's increased capacity and audio quality drove a surge in game sales, fueling industry growth.
Comment	The basic model response is verbose and includes peripheral historical details; TraceBias more concise focus on the primary mechanisms by which CD technology impacted gaming would enhance clarity and relevance.

Table 19: TopicalChat

Field	Content
Prompt	agent_B: "Did you know Steve Spurrier coach the Redskins?" ... agent_B:
label	"he has 12 victories in 32 games so maybe thats why he got fired , I think the team had high hopes on him."
Basic model	<p>It appears that agent_A and agent_B are discussing Steve Spurrier's career:</p> <ol style="list-style-type: none"> 1. Confirmed that Spurrier coached the Redskins. 2. Recounted anecdote about his duck-lips TV moment. 3. Noted his 1966 Heisman Trophy at Florida. 4. Identified misattribution of Tampa Bay Bandits. 5. Speculated on team embarrassment and website omission.
TraceBias	agent_B: "I know, its hard to find characters as known and famous as Steve lol..." agent_A: "I think you're right. I mean, Steve Spurrier ... need to give him the recognition he deserves."
Comment	The basic model submission fails to recognize the continuation task and instead provides a summary, whereas the TraceBias submission demonstrates proper dialogue continuation and contextual sensitivity.