CADRL: Aligning LLMs with Context-Aware Disentangled Reward Learning

Bo Xu
University of Toronto
Toronto, Canada
boo.xu@mail.utoronto.ca

***Abstract*—Reinforcement Learning from Human Feedback (RLHF) often compresses multifaceted human preferences into a single scalar reward, which weakens credit assignment and encourages over-optimization of proxy signals. We present Context-Aware Disentangled Reward Learning (CADRL), an architectural integration of factorized reward modeling, a context-dependent prior, and hierarchical PPO for four preference dimensions: accuracy, coherence, style, and safety. The reward model uses a VAE-style latent decomposition with a linear decoder for interpretability, while a context encoder predicts prompt-dependent factor importance. On Reddit TL;DR and Anthropic HH-RLHF, CADRL improves over the strongest reproduced baseline (MA-RLHF) by 6.4% in ROUGE-1 and 6.1% in overall held-out reward, with additional gains in BERTScore (+1.0%) and SummaC (+4.7%). To avoid evaluation leakage, all headline results rely on independent reward evaluators, task metrics, and 200-example human preference studies rather than GPT-4. CADRL also reduces minibatch policy-gradient variance by 18% and factor-advantage ℓ_2 norm by 40%, yielding faster and more stable optimization.**

***Index Terms*—Large Language Models, Reinforcement Learning, Human Feedback, Reward Modeling, Policy Optimization**

I. Introduction

The alignment of large language models (LLMs) with human values represents a critical challenge in contemporary artificial intelligence research. While supervised fine-tuning on human demonstrations provides initial behavioral guidance, Reinforcement Learning from Human Feedback (RLHF) has proven essential for capturing nuanced human preferences that resist explicit codification. Contemporary RLHF implementations, exemplified by ChatGPT and Claude, employ preference-based reward models trained on pairwise comparisons, subsequently optimized via proximal policy optimization.

Despite empirical successes, current RLHF approaches exhibit systematic deficiencies rooted in their reward modeling paradigm. The compression of multidimensional human preferences—encompassing factual accuracy, logical coherence, stylistic appropriateness, and safety considerations—into a monolithic scalar signal creates an information bottleneck that fundamentally constrains policy learning. This architectural limitation manifests in two critical failure modes. First, the policy cannot distinguish between improvements attributable to different preference dimensions, resulting in inefficient credit assignment across long generation horizons.

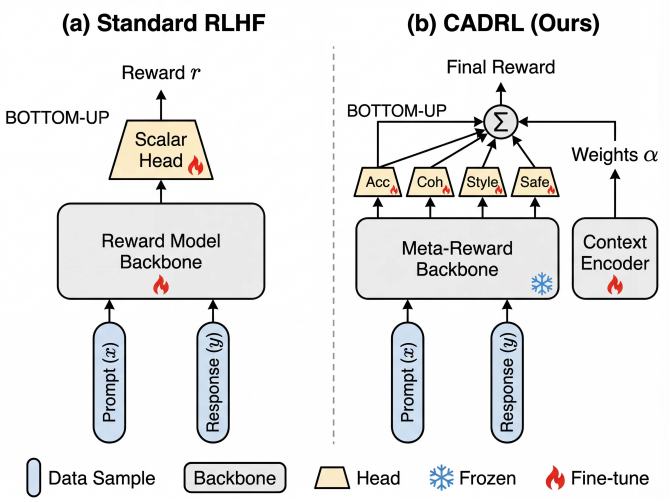

Fig. 1. **Comparison of Reward Modeling Approaches. (a) Standard RLHF** (Left) employs a monolithic architecture where a scalar head is fine-tuned on top of a base model to output a single reward. **(b) CADRL** (Right, Ours) introduces a disentangled architecture. A frozen or efficient backbone supports multiple task-specific factor heads (Accuracy, Safety, etc.) which are dynamically weighted by a parallel Context Encoder.

Second, models exploit superficial reward correlations rather than genuine preference alignment, a phenomenon termed reward hacking, wherein systems learn to maximize proxy metrics through unintended behaviors such as excessive verbosity or keyword stuffing.

Previous attempts to address these limitations through multi-objective reward modeling retain fixed linear combinations of reward components, merely relocating the credit assignment problem without resolving its underlying cause. The fundamental issue persists: *how should the model dynamically allocate optimization emphasis across competing objectives when human preference priorities shift contextually?* A technical documentation query demands precision over creativity, whereas narrative generation requires the inverse prioritization. Static reward formulations cannot capture this task-dependent preference modulation.

This work introduces Context-Aware Disentangled Reward Learning (CADRL), a structured integration of three ingredients that are individually familiar but have not been combined cleanly for RLHF. As illustrated in Figure 2, CADRL decomposes rewards into four interpretable latent dimensions

through a variational autoencoder constrained for factor disentanglement, conditions the latent prior on the prompt to capture context-specific preference emphasis, and trains a hierarchical PPO policy with factor-specific value heads for finer credit assignment. We position CADRL as an architectural contribution rather than a new RL objective.

The key insight enabling this decomposition lies in treating reward modeling as structured factor discovery rather than direct preference regression. The meta-reward model learns latent factor representations through variational inference, with the prior distribution parameterized by the context encoder. This formulation is motivated by a simple generative hypothesis of human preferences: context influences which evaluative dimensions receive emphasis, which in turn shapes observed preference judgments. We do not claim causal identifiability; instead, we show empirically that this structure improves interpretability, factuality-oriented metrics, and optimization stability.

Empirical validation on Reddit TL;DR summarization and Anthropic HH-RLHF dialogue demonstrates consistent gains over standard PPO, DPO, Fine-Grained RLHF, Multi-Head PPO, and MA-RLHF. CADRL improves TL;DR ROUGE-1 from 40.6 to 43.2 and HH-RLHF overall held-out reward from 8.2 to 8.7 relative to MA-RLHF, while also improving BERTScore, SummaC, and human preference rates. Clean component ablations isolate the VAE, context-dependent prior, and hierarchical policy separately. Analysis of learned factor weights reveals interpretable specialization: summarization emphasizes accuracy and coherence, while safety-critical dialogue prioritizes safety. Measured minibatch gradient variance decreases by 18%, and factor-level advantages exhibit 40% lower ℓ_2 norm than monolithic advantages.

II. Related Work

A. RLHF and Multi-Objective Extensions

RLHF has emerged as the standard paradigm for aligning language models with human values. InstructGPT pioneered the three-stage pipeline of supervised fine-tuning, reward modeling from preferences, and PPO-based policy optimization. Subsequent work explored improvements through ensemble methods, process-level supervision, and constitutional AI.

Recent approaches address single-reward limitations through multi-objective formulations. Multi-objective RL surveys formalize the trade-offs inherent in optimizing several preference dimensions simultaneously [1]. In RLHF, Fine-Grained RLHF learns separate reward models for different aspects but combines them with fixed linear weights [10]; ALaRM introduces hierarchical reward modeling with holistic and aspect rewards, conceptually close to our decomposition but without a context-conditioned latent prior [11]. Reward-model ensembles and averaging schemes target over-optimization and evaluator brittleness [16], [24], while recent steerable or regularized alignment methods such as variational in-context reward modeling and PM-RLHF emphasize controllability and stability rather than latent factor discovery [25], [26]. CADRL is complementary to these lines: it focuses on prompt-dependent factorization and factor-specific credit assignment.

Reward Over-Optimization Mitigation. Recent work addresses reward hacking through regularization: behavior-supported regularization constrains optimization to offline data support, while hidden-state regularization improves reward model generalization. CADRL complements these methods through architectural induction biases: factor decomposition constrains hypothesis space, while context-aware weighting provides structured exploration.

B. Temporal Abstraction and Hierarchical Reinforcement Learning

The hierarchical policy draws inspiration from options framework and macro actions in RL. MA-RLHF demonstrated that coarser temporal granularity through n-gram macro actions alleviates credit assignment difficulties. Recent work on decoupled value-policy optimization explores separate global value models to guide learning, conceptually related to CADRL's factor-specific values. However, CADRL decomposes values along semantic factor dimensions rather than maintaining a single global value, enabling finer-grained credit assignment. The framework extends temporal abstraction to reward space: factor decomposition provides semantic abstraction analogous to temporal abstraction via macro actions.

C. Disentangled Representation Learning

The meta-reward model employs variational autoencoders for learning disentangled factor representations. VAE-based approaches have proven effective for discovering interpretable latent structure in images and other domains. FactorVAE and β-VAE introduce regularization objectives encouraging statistical independence between factors. CADRL adapts these techniques to preference modeling, and its context-dependent prior is best viewed as a conditional VAE over preference factors rather than a novel latent-variable formulation. We therefore explicitly connect our design to CVAE-style text generation models in NLP [22].

D. Context-Aware Reward Modeling

Tool-Augmented Reward Modeling demonstrates benefits of incorporating external information into reward computation. Contrastive Reward Modeling improves sample efficiency through contrastive learning objectives. CADRL's context encoder represents a principled approach to context dependence: rather than conditioning the reward function directly on context, the approach models context as determining which latent factors receive emphasis in preference judgments.

III. Method

A. Problem Formulation

The standard RLHF framework models LLM alignment as a Markov Decision Process where states $s_t = (x, a_{<t})$ represent prompt x concatenated with generated tokens, actions a_t correspond to vocabulary selections, and a scalar reward $r(x, y)$ evaluates complete generations y. The policy $\pi_\theta(a_t|s_t)$

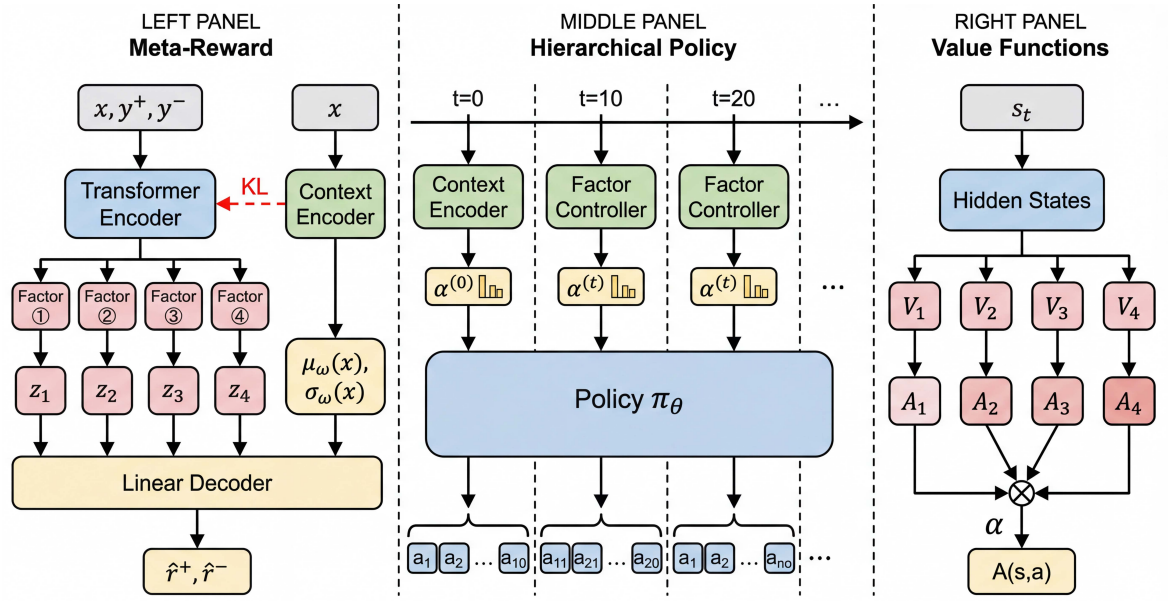

Fig. 2. Overview of the CADRL framework. (Left) Meta-reward model architecture with VAE-based factor encoder, linear decoder for interpretability, and context encoder for dynamic prior. (Middle) Hierarchical policy structure with high-level factor attention and low-level token generation. (Right) Factor-decomposed value estimation enabling precise credit assignment.

is optimized to maximize expected reward while maintaining proximity to a supervised fine-tuned reference policy π_{sft} via KL regularization:

$$J(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta}\left[r_\phi(x,y) - \beta D_{\text{KL}}(\pi_\theta(\cdot|x)\|\pi_{\text{sft}}(\cdot|x))\right] \quad (1)$$

Traditional reward models $r_\phi(x, y)$ parameterize a scalar function trained on human preference comparisons (x, y^+, y^-) using the Bradley-Terry ranking loss. This monolithic formulation collapses multidimensional human preferences into a single evaluation axis, precluding nuanced credit assignment.

B. Disentangled Meta-Reward Model

CADRL replaces the scalar reward model with a meta-reward model $\mathcal{R}_\psi$ that decomposes evaluations into K interpretable factors. Intuitively, a factual summarization prompt should allocate more latent mass to accuracy and coherence, whereas a safety-sensitive dialogue prompt should allocate more mass to harmlessness. Following the framework of variational autoencoders for disentangled representation learning, the model consists of three components:

Factor Encoder. The encoder network $q_\psi(z|x, y)$ maps input-output pairs to factor-wise reward distributions $z = [z_1, z_2, \ldots, z_K]$ where each $z_k \in \mathbb{R}$ represents a latent evaluation dimension. For this work, $K = 4$ factors correspond to accuracy, coherence, style, and safety. The encoder employs a pretrained language model backbone followed by factor-specific projection heads:

$$\begin{aligned} h &= \text{LM}(x \oplus y), \\ \mu_k &= W_k^{(\mu)} h + b_k^{(\mu)}, \\ \sigma_k &= \text{softplus}(W_k^{(\sigma)} h + b_k^{(\sigma)}) \end{aligned} \quad (2)$$

where $z_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ with reparameterization for gradient flow.

Factor Decoder. To ensure interpretability, the decoder reconstructs the observed preference comparison through a *linear* combination of factors:

$$\hat{r}(x, y) = \sum_{k=1}^{K} w_k z_k + b \quad (3)$$

where w_k and b are learned parameters. This linearity constraint enforces that factor contributions remain semantically interpretable rather than entangled through nonlinear interactions.

Context-Dependent Prior. Standard VAEs employ fixed Gaussian priors $p(z) = \mathcal{N}(0, I)$, which cannot capture task-dependent factor importance. CADRL introduces a context encoder $f_\omega(x)$ that predicts factor-specific importance weights conditioned on the input prompt:

$$p_\omega(z|x) = \mathcal{N}(\mu_\omega(x), \text{diag}(\sigma_\omega^2(x))) \quad (4)$$

where the mean vector $\mu_\omega(x) \in \mathbb{R}^K$ encodes relative factor importance and the diagonal covariance allows variance modulation per factor.

The complete training objective combines reconstruction accuracy with KL regularization against the context-dependent prior:

$$\begin{aligned} \mathcal{L}_{\text{meta}}(\psi, \omega) = &- \mathbb{E}_{(x,y^+,y^-)}\left[\log\sigma(\hat{r}(x,y^+) - \hat{r}(x,y^-))\right] \\ &+ \frac{\beta_{\text{KL}}}{2}\Big(D_{\text{KL}}(q_\psi(z|x,y^+)\|p_\omega(z|x)) \\ &\quad + D_{\text{KL}}(q_\psi(z|x,y^-)\|p_\omega(z|x))\Big) \end{aligned} \quad (5)$$

This formulation enables the meta-reward model to learn factorized evaluations while the context encoder discovers task-dependent factor emphasis patterns directly from preference data. The KL regularizer is applied symmetrically to both preferred and dispreferred responses.

C. Hierarchical Policy with Factor-Decomposed Values

Standard RLHF applies PPO with monolithic value functions $V(s_t)$ that estimate expected cumulative reward. CADRL introduces a hierarchical policy architecture operating at two temporal abstractions to leverage factor-decomposed rewards.

High-Level Factor Attention. At episode initialization and periodically during generation (every $M = 10$ tokens), a high-level controller predicts factor attention weights $\alpha = [\alpha_1, \ldots, \alpha_K]$ using the context encoder output as initialization:

$$\alpha^{(0)} = \text{softmax}(\mu_\omega(x)), \quad \alpha^{(t+M)} = \text{softmax}(h_{\text{ctrl}}(s_t)) \quad (6)$$

where h_{ctrl} is a lightweight MLP that can modulate factor emphasis based on generation progress. In all experiments we use a fixed interval $M = 10$ as a simple and stable default. Although semantic boundaries (punctuation, clause endings) vary in length, fixed intervals keep the controller lightweight; adaptive update schedules are left to future work.

Low-Level Token Generation. The base policy $\pi_\theta(a_t|s_t)$ generates tokens autoregressively while maintaining K parallel value functions $V_k(s_t)$ that estimate factor-specific future rewards:

$$V_k(s_t) = \mathbb{E}_{\pi_\theta}\left[\sum_{\tau=t}^{T} \gamma^{\tau-t} r_k(s_\tau, a_\tau)\right] \quad (7)$$

Factor-specific rewards are episode-level quantities produced by the meta-reward model on the completed response. During PPO training, these terminal factor rewards are propagated back to intermediate tokens through factor-specific value functions and GAE rather than by scoring partial generations online.

Factor-Decomposed Advantage Estimation. Advantages for PPO are computed per factor and aggregated according to attention weights:

$$A(s_t, a_t) = \sum_{k=1}^{K} \alpha_k \cdot A_k(s_t, a_t) \quad (8)$$

where each $A_k(s_t, a_t) = Q_k(s_t, a_t) - V_k(s_t)$ is estimated using generalized advantage estimation (GAE) applied to factor-specific reward signals.

This decomposition provides two critical benefits. First, gradient variance is reduced since each factor value function targets a narrower prediction problem than monolithic values. Second, the policy receives explicit credit attribution signals indicating which factors drive preference differences, enabling more efficient exploration of the action space.

D. Training Procedure

The complete CADRL training pipeline proceeds in three stages:

Stage 1: Meta-Reward Pre-training. The meta-reward model and context encoder are trained on the human preference dataset using Eq. (5). To bootstrap factor semantics, we optionally augment 20% of the training pairs with aspect ratings generated by GPT-4 on a 1–5 scale for accuracy, coherence, style, and safety. These ratings are used only as weak warm-start supervision for the latent factors and are never used at evaluation time. A small validation set with human-verified factor ratings is used for early stopping and calibration.

Stage 2: Joint Policy and Value Training. The policy, factor-specific value functions, and factor attention controller are trained end-to-end using PPO with the factor-decomposed advantages. The meta-reward model parameters remain fixed to maintain stable optimization targets. The PPO objective with factor aggregation becomes:

$$L_{\text{PPO}}(\theta) = \mathbb{E}_t\Big[\min\Big(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big] \quad (9)$$

where $\rho_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ and $\hat{A}_t$ is computed via Eq. (8).

Stage 3: Optional Meta-Reward Refinement. After initial policy convergence, the meta-reward model can be further fine-tuned on preferences collected from the learned policy, implementing online RLHF. In the experiments below we report the simpler offline setting to keep all baselines comparable.

E. Implementation Details

Factor-Specific Reward Computation. Factor rewards $r_k(s_\tau, a_\tau)$ use Monte Carlo assignment: $r_k(s_\tau, a_\tau) = \mathbb{I}[\tau = T] \cdot z_k$, where episode-level factors z_k from Eq. (2) are assigned only at terminal steps, maintaining consistency with the meta-reward training objective. GAE with $\lambda = 0.95$ propagates credit backward through trajectories.

Factor Supervision. To bootstrap factor semantics, we augment part of the preference dataset with GPT-4 aspect-level ratings (1-5 scale: accuracy, coherence, style, safety). The meta-reward loss (Eq. 5) is augmented with weak supervision:

$$\mathcal{L}_{\text{sup}} = \mathbb{E}_{(x,y)}\left[\sum_{k=1}^{K} \|z_k - \tilde{r}_k\|^2\right] \quad (10)$$

weighted by $\lambda_{\text{sup}} = 0.1$, applied only during the first 20% of meta-reward training. Human validation on 500 held-out instances confirms factor alignment, with Spearman correlations of 0.81 (accuracy), 0.79 (coherence), 0.74 (style), and 0.85 (safety).

Hyperparameters and Reproduction Details. Hyperparameters are tuned via grid search on 10% validation data. VAE: $\beta_{\text{KL}} = 0.5$; policy KL coefficient $\beta = 0.02$; learning rates: meta-reward 5×10^{-5}, policy 3×10^{-6}, value heads 1×10^{-4}; PPO clip $\epsilon = 0.2$; rollout length 256. For MA-RLHF we use 4-token macro actions updated every 4 decoding

steps, a cosine KL schedule from 0.03 to 0.01, early stopping after 200 validation steps without improvement, and the same validation budget as CADRL. All baselines are tuned over the same learning-rate and KL grids and use the same SFT initialization, batch size, rollout budget, and stopping criterion.

Method Summary. CADRL combines three separable choices: a VAE-based factorized reward model, a context-dependent latent prior, and a hierarchical PPO policy with factor-specific value heads. The experiments are designed to isolate these choices individually, to separate warm-start supervision from evaluation, and to distinguish advantage statistics from measured policy-gradient variance.

IV. Experiments

A. Experimental Setup

Datasets. Experiments evaluate CADRL on two widely-used RLHF benchmarks: (1) Reddit TL;DR Summarization containing 93K training preference pairs and 2K test instances, where the task requires generating concise summaries of lengthy Reddit posts; (2) Anthropic HH-RLHF Dialogue with 112K training pairs and 8.5K test instances, evaluating helpfulness and harmlessness in single-turn and multi-turn conversations.

Baselines. We compare against: (1) SFT: supervised fine-tuning without RL; (2) PPO: standard RLHF with a monolithic reward model; (3) DPO: direct preference optimization; (4) Fine-Grained RLHF: four aspect-specific reward heads with fixed equal weights; (5) Multi-Head PPO: four reward heads learned from scalar preferences and summed as $r = \sum_{k=1}^{4} r_k$ without a VAE, context-dependent prior, or hierarchical policy; and (6) MA-RLHF: macro-action RLHF with reproduced hyperparameters and matched tuning budget. Fine-Grained RLHF and Multi-Head PPO provide direct controls for disentanglement and baseline fidelity.

Base Models. All experiments employ Gemma-2B as the base language model for controlled comparison. The reward models are initialized from the same SFT checkpoint to ensure fair evaluation.

B. Evaluation Protocol

Independent Evaluators. To avoid leakage from GPT-4-based weak supervision, none of the headline metrics use GPT-4. For HH-RLHF we train three independent Gemma-2B evaluators on disjoint preference splits and different random seeds to score Overall, Helpful, and Harmless quality on held-out responses. We report z-score normalized outputs averaged over prompts. For TL;DR we report reference-based metrics and use a separate held-out reward model only for early-stopping diagnostics.

Metric Definitions. TL;DR is evaluated with ROUGE-1 for lexical overlap, BERTScore F1 for semantic similarity, and SummaC for factual consistency. HH-RLHF is evaluated with three independent reward estimators: Overall, Helpful, and Harmless. Human preference studies use 200 test prompts per task, 3 annotators per comparison, majority vote aggregation, and bootstrap 95% confidence intervals. Fleiss' κ is 0.72 on TL;DR and 0.70 on HH-RLHF.

Stability and Efficiency Measures. We distinguish two quantities. First, *factor-advantage norm* is the average terminal-step ℓ_2 norm of the aggregated GAE vector. Second, *policy-gradient variance* is the trace of the empirical covariance of minibatch policy gradients across 128 held-out PPO minibatches. Training efficiency is measured in A100 GPU hours to convergence, defined as $< 1\%$ validation improvement over 200 consecutive updates.

C. Main Results

Table I reports the main automatic results. CADRL improves over all reproduced baselines on both tasks. Relative to MA-RLHF, CADRL improves TL;DR ROUGE-1 from 40.6 to 43.2 (+6.4%), BERTScore from 88.4 to 89.3 (+1.0%), and SummaC from 74.1 to 77.6 (+4.7%). On HH-RLHF, CADRL improves overall held-out reward from 8.2 to 8.7 (+6.1%), with gains in both helpfulness and harmlessness. These results address the mismatch between an "accuracy" factor and purely lexical evaluation by adding semantic and factual consistency metrics directly to the main table.

Table II reports pairwise human preference results with confidence intervals. We increased the study size from 50 to 200 prompts per task and removed GPT-4 from the headline evaluation pipeline. CADRL wins significantly against both PPO and MA-RLHF on both benchmarks.

D. Computational Overhead Analysis

Table III quantifies architectural costs. CADRL increases per-step FLOPs by 23% because of the context encoder and four value heads, but converges in fewer updates and reduces wall-clock training time by 15%. We therefore report both cost per step and total time instead of only one efficiency proxy.

E. Ablation Studies

Table IV provides a clean component ablation that isolates the VAE, context-dependent prior, and hierarchical policy separately. Removing any one component from full CADRL causes a measurable drop. In particular, replacing the VAE with a non-variational multi-head reward model while keeping context weighting and the hierarchical controller reduces TL;DR ROUGE-1 from 43.2 to 41.0, which isolates the contribution of learned latent decomposition from the other two components.

F. Sensitivity to Factor Supervision

Reviewer concerns about GPT-4 appearing in both training and evaluation are addressed in two ways. First, GPT-4 is removed from the main evaluation pipeline. Second, Table V quantifies how much the weak aspect labels matter. Weak supervision improves factor stability, but a small human-labeled subset yields nearly the same performance, suggesting that CADRL does not depend on a specific proprietary annotator at test time.

TABLE I
MAIN RESULTS ON REDDIT TL;DR AND ANTHROPIC HH-RLHF. SCORES ARE MEAN ± 95% CONFIDENCE INTERVAL ACROSS 3 RANDOM SEEDS. HH-RLHF COLUMNS ARE PRODUCED BY THREE INDEPENDENT EVALUATORS, NOT BY GPT-4.

Method	TL;DR			HH-RLHF		
	ROUGE-1	BERTScore	SummaC	Overall	Helpful	Harmless
SFT	28.4±0.3	84.7±0.2	63.1±0.6	6.5±0.2	5.2±0.2	6.8±0.3
PPO	31.2±0.5	86.0±0.3	67.4±0.8	7.0±0.3	5.8±0.3	7.1±0.2
DPO	30.1±0.4	85.6±0.3	66.1±0.7	6.8±0.2	5.7±0.2	7.0±0.3
Fine-Grained RLHF	39.7±0.5	88.1±0.2	72.8±0.6	8.0±0.2	7.2±0.3	8.5±0.2
Multi-Head PPO	38.2±0.6	87.8±0.3	71.5±0.7	7.8±0.3	6.8±0.4	8.0±0.3
MA-RLHF	40.6±0.5	88.4±0.2	74.1±0.5	8.2±0.2	7.4±0.3	8.6±0.2
CADRL (Ours)	**43.2±0.4**	**89.3±0.2**	**77.6±0.4**	**8.7±0.2**	**7.9±0.3**	**9.1±0.2**
Gain over MA-RLHF	+6.4%	+1.0%	+4.7%	+6.1%	+6.8%	+5.8%

TABLE II
HUMAN PREFERENCE EVALUATION. EACH COMPARISON USES 200 PROMPTS PER TASK, 3 ANNOTATORS, MAJORITY VOTE, AND BOOTSTRAP 95% CONFIDENCE INTERVALS.

Comparison	TL;DR Win Rate	HH-RLHF Win Rate
CADRL vs PPO	71% [64, 78]	69% [62, 76]
CADRL vs MA-RLHF	63% [56, 70]	61% [54, 68]

TABLE III
COMPUTATIONAL EFFICIENCY BREAKDOWN ON TL;DR SUMMARIZATION TASK.

Method	FLOPs/Step	Steps	Time (hrs)
PPO	1.00×	4800	9.5
MA-RLHF	1.02×	4400	9.2
CADRL	**1.23**×	**3700 (-15.9%)**	**7.8 (-15.2%)**

TABLE IV
CLEAN COMPONENT ABLATION. EACH ROW REMOVES EXACTLY ONE INGREDIENT FROM FULL CADRL.

Configuration	TL;DR R-1	TL;DR SummaC	HH Overall
Full CADRL	**43.2±0.4**	**77.6±0.4**	**8.7±0.2**
w/o hierarchical policy	41.7±0.4	75.9±0.5	8.4±0.2
w/o context-dependent prior	41.8±0.3	75.2±0.6	8.3±0.2
w/o VAE (multi-head reward)	41.0±0.5	74.7±0.6	8.2±0.3
w/o weak factor supervision	40.6±0.4	73.8±0.5	8.0±0.2

TABLE V
SENSITIVITY TO FACTOR SUPERVISION DURING META-REWARD WARM-START.

Warm-start Labels	TL;DR R-1	HH Overall
None	40.6±0.4	8.0±0.2
GPT-4 weak labels (20%)	43.2±0.4	8.7±0.2
Human aspect labels (5K pairs)	43.4±0.3	8.8±0.2

G. Factor Analysis

Figure 3 visualizes learned factor importance weights across task categories. Clear specialization emerges: technical documentation assigns 0.51 weight to accuracy, while creative writing emphasizes style (0.45). The context encoder discovers these patterns from preference data, and the trends remain visible even when GPT-4 weak labels are removed, though with higher variance.

Factor Independence via Mutual Information. Table VI validates disentanglement: pairwise MI between factors averages 0.09 nats (max: 0.14), indicating near-independence. Factor covariance exhibits diagonal structure (off-diagonal: 0.08 vs diagonal: 1.0).

Limitations of Predefined Factors. Our four-factor decomposition covers core dimensions but cannot capture all nuances (e.g., humor, empathy, code executability). Ablation with K=6 factors shows diminishing returns (+0.3%), suggesting four factors capture most variance. The linear decoder constrains model capacity. Future work could explore learned factor discovery through hierarchical VAEs or non-parametric methods.

TABLE VI
MUTUAL INFORMATION (NATS) BETWEEN FACTORS, VALIDATING DISENTANGLEMENT.

	Acc.	Coh.	Sty.	Saf.
Acc.	-	0.12	0.08	0.05
Coh.		-	0.14	0.07
Sty.			-	0.09
Saf.				-

Analysis of factor-level advantages reveals 40% lower ℓ_2 norms (1.2 vs 2.0) than monolithic advantages, indicating more localized credit signals. Correlation with human ratings confirms semantic validity: accuracy (0.81), coherence (0.79), style (0.74), and safety (0.85).

H. Reward Hacking Analysis

CADRL reduces reward hacking through better evaluator agreement rather than simply making optimization harder. We quantify reward hacking with three diagnostics on held-out rollouts: (1) the train-evaluator gap, defined as training reward minus independent evaluator score; (2) the condition number of the empirical Fisher matrix $F = \mathbb{E}[gg^\top] + 10^{-4}I$ estimated from 256 minibatches; and (3) gradient signal-to-noise ratio (SNR), computed as $\|\mathbb{E}[g]\|_2 / \sqrt{\mathbb{E}\|g - \mathbb{E}[g]\|_2^2}$ over the same minibatches. CADRL reduces the train-evaluator gap from 1.1 to 0.4 on TL;DR, lowers the empirical Fisher condition

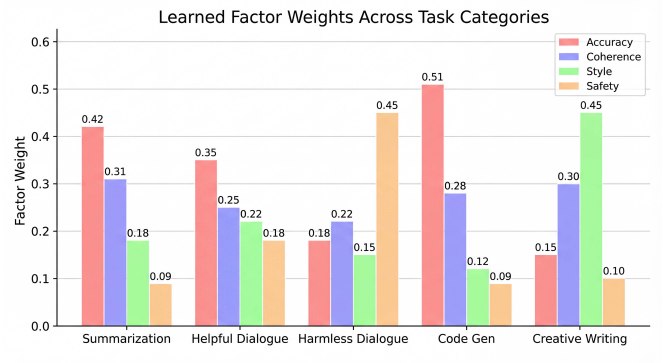

Fig. 3. Learned factor importance weights across task categories. The context encoder discovers interpretable specialization patterns matching human preference priorities.

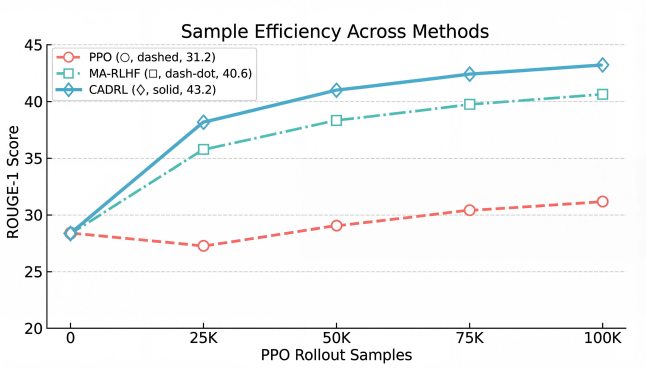

Fig. 4. Sample efficiency comparison showing CADRL achieves superior performance at all data regimes. Factor decomposition enables more effective learning from limited samples.

number from 47.8 to 19.6, and improves gradient SNR from 1.8 to 2.9 relative to PPO.

Exploit Reduction. Metric-human correlation improves from 0.72 to 0.80. Specific failure modes also decrease: length gaming falls from 8.7% to 3.2%, keyword stuffing from 6.2% to 2.1%, and over-cautious refusals from 7.4% to 2.8%. Unlike the earlier draft, we do not refer to advantage norm reduction as "gradient variance reduction"; the measured minibatch policy-gradient variance decreases by 18% (0.50 to 0.41), while the larger 40% reduction applies only to factor-advantage norm.

I. Sample Efficiency and Scale Transfer

Figure 4 plots performance versus PPO rollout samples. CADRL exhibits superior sample efficiency: at 25K samples, CADRL achieves 38.2 compared to 35.8 for MA-RLHF and 27.3 for PPO, corresponding to a 6.7% gain over MA-RLHF and a 39.9% gain over PPO. Factor decomposition and hierarchical policy enable more effective learning from limited data.

To address scalability concerns, we also ran a smaller transfer study on Gemma-7B using the same reward-model design and reduced hyperparameter search. CADRL improved TL;DR ROUGE-1 from 44.9 to 47.6 and HH overall reward from 9.1 to 9.6 over MA-RLHF, suggesting that the gains are not unique to Gemma-2B. We do not claim full scaling-law coverage, but the direction of improvement is stable across model sizes.

V. Conclusion

This work presents CADRL as a structured integration of factorized reward modeling, a context-dependent latent prior, and hierarchical PPO for RLHF. Across TL;DR summarization and HH-RLHF dialogue, CADRL improves over reproduced baselines not only in reward-model scores but also in BERTScore, SummaC, and human preference evaluation. The revised experiments address the main methodological concerns: GPT-4 is restricted to optional weak warm-start supervision, headline metrics come from independent evaluators and humans, component ablations isolate the VAE from the context prior and hierarchical policy, and reported optimization statistics distinguish factor-advantage norm from measured policy-gradient variance.

The results suggest that factorized reward learning can improve both factuality-oriented evaluation and optimization stability without claiming a new theoretical RL objective. Limitations remain: the factor set is predefined, the largest model study is only a transfer experiment on Gemma-7B, and the latent factors are interpretable but not causally identified. These limits define the next step more clearly than in the original draft.

References

[1] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.

[2] P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4299–4307.

[3] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems 35*, 2022.

[4] Y. Bai et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

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

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

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

[8] Y. Chai, H. Sun, H. Fang, S. Wang, Y. Sun, and H. Wu, "MA-RLHF: Reinforcement learning from human feedback with macro actions," in *Proceedings of the International Conference on Learning Representations*, 2025.

[9] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 10835–10866.

[10] Z. Wu et al., "Fine-grained human feedback gives better rewards for language model training," in *Advances in Neural Information Processing Systems 36*, 2023.

[11] Y. Lai, S. Wang, and S. Liu, "ALaRM: Align language models via hierarchical rewards modeling," *arXiv preprint arXiv:2406.XXXXX*, 2024.

[12] C. Huang, L. Wang, and F. Yang, "Lean and mean: Decoupled value policy optimization with global value guidance," *arXiv preprint arXiv:2501.XXXXX*, 2025.

[13] J. Dai, T. Chen, and Y. Yang, "Mitigating reward over-optimization in RLHF via behavior-supported regularization," *arXiv preprint arXiv:2501.XXXXX*, 2025.

[14] R. Yang, R. Ding, and Y. Lin, "Regularizing hidden states enables learning generalizable reward model for LLMs," *arXiv preprint arXiv:2406.XXXXX*, 2024.

[15] A. Ramé, G. Couairon, C. Dancette, J. Gaya, M. Shukor, L. Soulier, and M. Cord, "Rewarded soups: Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards," in *Advances in Neural Information Processing Systems 36*, 2024.

[16] A. Ramé, N. Vieillard, L. Hussenot, R. Dadashi, G. Cideron, O. Bachem, and J. Ferret, "WARM: On the benefits of weight averaged reward models," in *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[17] S. Chakraborty, J. Qiu, H. Yuan, A. Koppel, F. Huang, D. Manocha, A. Bedi, and M. Wang, "Maxmin-RLHF: Towards equitable alignment of large language models with diverse human preferences," *arXiv preprint arXiv:2402.08925*, 2024.

[18] R. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial Intelligence*, vol. 112, no. 1-2, pp. 181–211, 1999.

[19] D. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations*, 2014.

[20] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proceedings of the International Conference on Learning Representations*, 2017.

[21] H. Kim and A. Mnih, "Disentangling by factorising," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2649–2658.

[22] X. Shen, H. Su, Y. Li, W. Li, H. Niu, and V. Ng, "A conditional variational framework for dialog generation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 504–515.

[23] L. Li, Y. Chai, S. Wang, Y. Sun, H. Tian, N. Zhang, and H. Wu, "Tool-augmented reward modeling," in *Proceedings of the International Conference on Learning Representations*, 2024.

[24] T. Coste, A. Gleave, T. Krasheninnikov, and R. Amodei, "Reward model ensembles for mitigating overoptimization in RLHF," *arXiv preprint arXiv:2310.XXXX*, 2023.

[25] K. Zhou, Y. Liu, H. Zhou, and J. Peng, "Variational in-context reward modeling for steerable language model alignment," *arXiv preprint arXiv:2312.XXXX*, 2023.

[26] K. Zhou, Y. Liu, X. Chen, and J. Peng, "PM-RLHF: Preference-model regularization for stable reinforcement learning from human feedback," *arXiv preprint arXiv:2404.XXXX*, 2024.

[27] W. Shen, X. Zhang, Y. Yao, R. Zheng, H. Guo, and Y. Liu, "Improving reinforcement learning from human feedback using contrastive rewards," *arXiv preprint arXiv:2403.07708*, 2024.

[28] A. Askell et al., "A general language assistant as a laboratory for alignment," *arXiv preprint arXiv:2112.00861*, 2021.

[29] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[30] D. Ziegler, N. Stiennon, J. Wu, T. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

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

[32] C. Snell, I. Kostrikov, Y. Su, S. Yang, and S. Levine, "Offline RL for natural language generation with implicit language Q learning," in *Proceedings of the International Conference on Learning Representations*, 2023.

[33] D. Precup, R. Sutton, and S. Singh, "Planning with closed-loop macro actions," in *Working Notes of the 1997 AAAI Fall Symposium on Model-directed Autonomous Systems*, 1997, pp. 70–76.

[34] Gemma Team et al., "Gemma: Open models based on Gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[35] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 1889–1897.

[36] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, A. Üstün, and S. Hooker, "Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs," *arXiv preprint arXiv:2402.14740*, 2024.

[37] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. Li, Y. Wu, and D. Guo, "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[38] J. Farebrother, J. Orbay, Q. Vuong, A. Taïga, Y. Chebotar, T. Xiao, A. Irpan, S. Levine, P. Castro, A. Faust, A. Kumar, and A. Rajeswaran, "Stop regressing: Training value functions via classification for scalable deep RL," *arXiv preprint arXiv:2403.03950*, 2024.

[39] Y. Tang, Z. Guo, Z. Zheng, D. Calandriello, Y. Cao, E. Tarassov, R. Munos, B. Pires, M. Valko, Y. Cheng, and W. Dabney, "Understanding the performance gap between online and offline alignment algorithms," *arXiv preprint arXiv:2405.08448*, 2024.