

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 GROUP-NORMALIZED IMPLICIT VALUE OPTIMIZATION FOR LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Fine-tuning Large Language Models (LLMs) with reinforcement learning (RL) has become a key technique for enhancing performance on a wide range of tasks, from user alignment to complex reasoning. However, this approach is often hindered by the difficulty of fine-grained credit assignment, as it typically relies on sparse rewards given only at the end of a completely generated sequence. Conventional solutions often require training an auxiliary value network known as critic, which introduces significant computational overhead and training instability. We present *Group-Normalized Implicit Value Optimization (GN-IVO)*, a novel, critic-free algorithm that directly addresses this challenge. GN-IVO learns step-level values implicitly from the policy through a group-normalized distributional matching objective. This approach elegantly circumvents the need for an explicit critic and avoids the computation of the intractable partition function by normalizing values across a group of sampled model responses. Theoretically, we prove that our objective recovers the true value function up to a constant, guaranteeing that the optimal policy is preserved. We demonstrate the practical effectiveness of GN-IVO on a diverse set of text generation and reasoning tasks, showing that it consistently outperforms strong RL baselines for LLMs.

1 INTRODUCTION

Reinforcement Learning (RL) has emerged as a powerful paradigm for post-training Large Language Models (LLMs) (Ouyang et al., 2022; Kumar et al., 2025). Its success is well-established across a range of critical applications, from aligning model outputs with human preferences (Bai et al., 2022; Rafailov et al., 2023) to enhancing their complex, multi-step reasoning abilities (Zelikman et al., 2024; DeepSeek-AI, 2025). Yet, many RL approaches for LLMs (Rafailov et al., 2023; Ethayarajh et al., 2024; Ahmadian et al., 2024) cast policy learning as a contextual bandit, where the entire generated response is treated as a single action that is assigned a scalar score by a reward model. This formulation clashes with the inherently sequential, long-horizon nature of both token-by-token generation and step-by-step reasoning, in which intermediate decisions compound over time. Moreover, outcome-only rewards yield a sparse and delayed learning signal, offering little guidance about which token or step-level decisions drive the final quality. As a result, step-level credit assignment remains opaque, suggesting that language generation should be treated as sequential decision making rather than a one-shot bandit.

Recent works (Zeng et al., 2024; Wang et al., 2024; 2025) therefore model LLM generation as a Markov Decision Process (MDP) to incorporate the sequential decision making. A common approach is to enforce Bellman consistency and learn step-aware value estimates. In particular, Wang et al. (2025); Liu et al. (2024) train policies within the soft Q-learning framework (Haarnoja et al., 2017), parameterizing the policy as a Boltzmann distribution over Q-values with a partition function. However, because the partition function must still be estimated, these methods retain an auxiliary estimator. While this estimator can serve as a baseline for variance reduction in policy updates, this introduces practical overhead; additional networks to train and tune, increased computational and memory costs, and added complexity in the post-training pipeline. This trade-off, gaining step-aware feedback at the expense of a more complex training procedure, motivates methods that enable effective sequential learning without auxiliary networks or substantial engineering burden.

054 In this work, we propose *Group-Normalized Implicit Value Optimization (GN-IVO)*, an RL fine-
 055 tuning method that incorporates the sequential nature of language generation without requiring any
 056 auxiliary network. We first generalize the canonical KL-regularized objective for LLMs to partial
 057 sequence, deriving an explicit link between the policy and a soft value function. This value function
 058 quantifies the contribution of a partial sequence to the eventual outcome reward. To learn these
 059 values, we introduce a *group-normalized* objective. For a given query, we sample a group of complete
 060 responses and form a target distribution over them, proportional to their exponential rewards. We
 061 then match this target with the normalized model’s predicted values for the partial sequence. This
 062 distributional matching forces the model to align its value predictions with the ground-truth relative
 063 values. We prove that this objective recovers the true value function up to an additive constant, which
 064 preserves the action preferences needed for the optimal policy. Finally, inspired by recent advances
 065 like DPO (Rafailov et al., 2023) that bypass explicit reward models, we likewise bypass explicit value
 066 function modeling by the direct policy-value relationship.
 067

068 The result is a lightweight and direct optimization scheme that avoids computing the intractable
 069 partition function: normalization within each sampled group cancels any additive constant including
 070 the partition function. Thus, our method provides fine-grained feedback on partial sequences without
 071 incurring the overhead of training a separate value estimator. We validate the proposed approach on a
 072 diverse set of text generation and reasoning tasks, including the Helpful assistant, TL;DR summarization,
 073 prompt generation for text-to-image models tasks, and challenging reasoning benchmarks. Across these scenarios, GN-IVO demonstrates its effectiveness in performance compared to standard
 074 RL fine-tuning methods for LLMs.
 075

We summarize our main contributions as follows:

- 076 **1. New Fine-Tuning Objective:** We introduce Group-Normalized Implicit Value Optimization, a
 077 novel RL objective that incorporates step-level feedback into LLM post-training. Our method
 078 dispenses with any auxiliary value/critic model, instead using group-based normalization to infer
 079 the relative value of partial sequences.
- 080 **2. Theoretical Guarantee:** We present a theoretical analysis showing that our normalized objective
 081 learns the true value function up to a constant offset, which does not affect the optimal policy.
 082 This result guarantees that optimizing our objective yields the same optimal policy as the original
 083 RL problem, thereby ensuring policy optimality and consistency.
- 084 **3. Empirical Validation:** We demonstrate the effectiveness of our approach on multiple benchmarks,
 085 including text generation and complex reasoning tasks. Our experiments show that the proposed
 086 method achieves higher rewards than standard RL algorithms, highlighting its practical benefit in
 087 aligning LLMs with sparse feedback.

089 2 RELATED WORKS

091 **Reinforcement Learning from Human Feedback (RLHF)** LLMs are often aligned with human
 092 preferences using RLHF, wherein a reward model trained on human feedback guides the policy
 093 optimization. PPO (Schulman et al., 2017) has become the canonical algorithm in early large-scale
 094 alignment works (Ouyang et al., 2022). Despite its success, PPO-based RLHF can be resource-
 095 intensive and operationally complex. It typically requires keeping multiple models in memory, a
 096 policy, a frozen reference model for KL control, a critic model, and a reward model. The performance
 097 is also sensitive to hyperparameters due to high variance of the policy gradient estimates and
 098 delayed outcome rewards. These challenges have motivated research into simpler fine-tuning methods
 099 that bypass the full RL machinery while aiming to preserve alignment quality. DPO (Rafailov
 100 et al., 2023) reframes the KL-regularized objective as a supervised loss on preference-labeled data,
 101 eliminating explicit reward model training, and thereby improving stability and computational
 102 efficiency. Numerous extensions build on this paradigm (Maeng et al., 2025; Qi et al., 2024). Distinct
 103 from these preference-based methods, DRO (Richemond et al., 2024) addresses scenarios where
 104 feedback is provided as a score for each completion, rather than as a pairwise preference. To handle
 105 this format, DRO incorporates the Soft Bellman equation into a bandit-setting objective, effectively
 106 learning from direct reward labels. **BRAIN** (Pandey et al., 2024) formulates RLHF as a distribution-
 107 matching problem between the target and learned policies, proposing a self-normalized baseline to
 108 reduce variance. However, DPO, DRO, **BRAIN** and most of their variants formulate the generation
 109 task as a bandit problem where generating a completion is considered a single action that receives a

108 single reward. This formulation obscures the sequential structure of language generation, thereby
 109 limiting performance on long-horizon tasks such as mathematical reasoning (Yuan et al., 2025).
 110

111 **Reinforcement Learning for LLM Reasoning** Beyond human preference alignment, recent works
 112 have explored RL to improve LLMs’ reasoning abilities, which require generating long chains of
 113 intermediate steps. Such training imposes a substantial memory demand due to long generated
 114 trajectories, even when the base model is modest in size. Since the critic network adds to this
 115 memory cost, a key research direction has been to develop approaches that remove explicit value
 116 function learning in policy-gradient methods. For instance, GRPO (Shao et al., 2024) is a variant of
 117 PPO that dispenses with a learned critic and instead computes advantages from group-wise return
 118 statistics within each batch. By using the average return from a group of trajectories as a baseline,
 119 GRPO significantly reduces memory overhead while maintaining training stability. In a similar vein,
 120 RLOO (Ahmadian et al., 2024) extends REINFORCE with a leave-one-out baseline. It computes
 121 the advantage of each response by comparing its reward to the average reward of others in the
 122 group, thereby reducing variance without requiring a separate value predictor. Despite their improved
 123 memory-efficiency, methods like GRPO and RLOO still treat the entire generated sequence as a
 124 single action. By doing so, they fail to model the internal structure of multi-step reasoning and assign
 125 credit equally across all tokens based on a single, final reward. This simplistic treatment overlooks the
 126 chain-of-thought or stepwise dependencies within complex reasoning processes, making it difficult to
 127 pinpoint which parts of the generation led to a successful outcome. Recently, OREO (Wang et al.,
 128 2025) and DQO (Liu et al., 2024) have formulated reasoning steps as an MDP, employing soft
 129 Q-learning to learn the policy. However, the Bellman error minimization requires an additional value
 130 network, making the training procedure complex and memory-inefficient, which hinders scalability.
 131 One the other hand, VinePPO (Kazemnejad et al., 2025) performs credit assignment without a separate
 132 network by conducting multiple Monte Carlo rollouts at every step. While effective in improving
 133 performance, this approach incurs a substantial sampling overhead.

134 To bridge this gap, we introduce Group-Normalized Implicit Value Optimization (GN-IVO), a method
 135 designed to be both sequential and critic-free. Similar to GRPO and RLOO, GN-IVO utilizes a group
 136 of K samples to remain critic-free. However, a key difference lies in how these samples are used:
 137 whereas GRPO computes advantages using batch statistics, GN-IVO uses the samples to form a group
 138 distributional matching objective. At the same time, and in spirit of OREO and DQO, our method
 139 explicitly models the step-by-step nature of generation to enable fine-grained credit assignment.

3 METHOD

3.1 BACKGROUND: KL-REGULARIZED POLICY OPTIMIZATION

141 KL-regularized policy optimization is a widely used approach for refining language model policies.
 142 The goal is to train a policy π_θ that maximizes the expected reward while staying close to a reference
 143 policy $\pi_{\theta_{\text{old}}}$, which is often the supervised fine-tuned (SFT) model that we start with. This objective
 144 is formalized as:

$$145 \max_{\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(\cdot|x)} \left[R(x, y) - \alpha \log \frac{\pi_{\theta}(y|x)}{\pi_{\theta_{\text{old}}}(y|x)} \right] \quad (1)$$

146 where x is a query sampled from the dataset \mathcal{D} , and y is a completion generated by the policy π_θ .
 147 The function $R(x, y)$ assigns a scalar reward to the completion y for a given query x , and $\alpha \geq 0$ is a
 148 temperature coefficient. The optimal policy π_{θ^*} for Eq. 1 has a known closed-form solution:

$$149 \pi_{\theta^*}(y|x) = \frac{\pi_{\theta_{\text{old}}}(y|x) e^{R(x,y)/\alpha}}{Z(x)} \quad (2)$$

150 where $Z(x) = \mathbb{E}_{\pi_{\theta_{\text{old}}}(y|x)}[e^{R(x,y)/\alpha}]$ serves as the partition function to ensure the distribution is
 151 normalized. While the formulation in Eq. 1 is standard for RL post-training of LLMs, it relies on a
 152 single reward $R(x, y)$ assigned to the entire generated sequence y .

153 However, this single reward for the entire sequence provides a weak credit assignment signal. Given
 154 the autoregressive, token-by-token generation process of language models, a more granular credit-
 155 assignment scheme is warranted. We therefore reformulate the objective to incorporate expected
 156 rewards at each timestep, enabling denser supervision.

162 3.2 LEARNING STEP-LEVEL VALUES VIA GROUP NORMALIZATION
163

164 Our first key theoretical result establishes that the relationship in Eq. 2 extends to arbitrary sequence
165 prefixes. Let $y = (y_0, \dots, y_{T-1})$ be any sequence of length T , where y_t denotes the token or
166 reasoning step at timestep t .¹

167 **Theorem 3.1.** *Suppose that π_{θ^*} and $\pi_{\theta_{\text{old}}}$ are autoregressive policies that generate sequences token-
168 by-token, and that satisfy Eq. 2 for complete response y . For any $t \in \{1, \dots, T\}$, let $y_{<t} := y_{0:t}$
169 denote the length- t prefix. Then, $y_{<t}$ distributions of these policies satisfy:*

$$170 \quad \pi_{\theta^*}(y_{<t}|x) = \frac{\pi_{\theta_{\text{old}}}(y_{<t}|x)e^{V(x, y_{<t})}}{Z(x)} \quad (3)$$

172 where a soft value function $V(x, y_{<t})$, which represents the expected future reward from $y_{<t}$, is
173 defined as:

$$174 \quad V(x, y_{<t}) := \begin{cases} R(x, y)/\alpha & t = T, \\ \log \mathbb{E}_{\pi_{\theta_{\text{old}}}(y|y_{<t}, x)}[e^{R(x, y)/\alpha}] & t < T. \end{cases} \quad (4)$$

177 While distributions over partial sequences are typically hard to analyze, Theorem 3.1 reveals an
178 exact analytical form for $\pi_{\theta^*}(y_{<t}|x)$ with $V(x, y_{<t})$. The value function $V(x, y_{<t})$ provides a formal
179 solution to the temporal credit assignment problem by quantifying the expected achievable reward
180 given the tokens generated so far, $y_{<t}$. It allows the model to attribute the final monolithic reward
181 back to each intermediate decision, enabling granular policy optimization.

183 **The challenge of evaluating the partition function** While the soft value function from Theorem 3.1
184 can be modeled with an auxiliary network V_ψ , doing so adds complexity and computational cost.
185 Inspired by DPO (Rafailov et al., 2023) that bypasses an explicit reward model, we instead pursue
186 a direct policy objective that avoids explicit value modeling. By rearranging Eq. 3, we can express
187 the value function implicitly in terms of a policy ratio, which allows us to reframe value fitting as a
188 policy optimization problem:

$$189 \quad V(x, y_{<t}) = \log Z(x) + \log \frac{\pi_{\theta^*}(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)}. \quad (5)$$

191 A natural starting point for policy training loss is to minimize the mean-squared error (MSE) between
192 the policy-defined value (Eq. 5) and the ground-truth value function (Eq. 4). Substituting our trainable
193 policy π_θ for the unknown optimal policy π_{θ^*} yields the following loss function:

$$194 \quad \mathcal{L}_{\text{MSE}}(Z, \theta) = \left(\log Z(x) + \log \frac{\pi_\theta(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)} - \log \mathbb{E}_{y \sim \pi_{\theta_{\text{old}}}(\cdot|y_{<t}, x)}[e^{R(x, y)/\alpha}] \right)^2. \quad (6)$$

197 A practical difficulty in Eq. 6, however, is the presence of the partition function $Z(x)$. Its computation
198 requires marginalizing over all possible sequences under $\pi_{\theta_{\text{old}}}$, which is intractable. While this term
199 can be approximated with an auxiliary network, $Z_\phi(x)$, doing so introduces additional complexity
200 and potential instability. This motivates a different approach that circumvents the need to estimate the
201 partition function entirely.

202 **The group-normalized objective** To avoid an auxiliary network for the partition function, we
203 propose a *group-normalized* objective for value estimation. The key insight is that for a fixed query
204 x , the partition function $Z(x)$ is a constant across all candidate responses. By normalizing values
205 over a sampled group of candidates, any shared constant including $\log Z(x)$ cancels out, making its
206 explicit estimation unnecessary. We first develop this objective for an explicit estimator, $V_\psi(\cdot)$, before
207 extending it to implicit, policy-only training.

208 The procedure starts by sampling a group of $K > 1$ i.i.d. completions, $\{y^{(i)}\}_{i=0}^{K-1}$, from $\pi_{\theta_{\text{old}}}(\cdot|x)$.
209 For a set of prefixes $\{y_{<t}^{(i)}\}_{i=0}^{K-1}$ at a randomly sampled timestep $t \leq T^2$, we propose the training

210 ¹We consider two types of y_t . For general text generation, y_t is a single token, $y_t = w_t \in \mathcal{V}$. For reasoning
211 tasks, y_t denotes a complete reasoning step represented as a token subsequence $y_t = (w_{t_0}, \dots, w_{t_{l_t-1}})$ with
212 $l_t \geq 1$.

213 ²We assume all K completions have the same length T after right-padding with [PAD] tokens. This is a
214 mild assumption, as sequences are typically padded to a fixed maximum length in practice. Both the policy and
215 reward models are designed to ignore padded positions, ensuring rewards are unaffected.

216 objective given by:
 217

$$218 \min_{V_\psi} \mathbb{E}_{\substack{x \sim \mathcal{D}, t \sim \mathcal{U}\{1, \dots, T\}, \\ y^{(0:K)}, y_{<t}^{(0:K)} \sim \pi_{\theta_{\text{old}}}(\cdot|x)}} \left[- \sum_{i=0}^{K-1} e^{R(x, y_{<t}^{(i)})/\alpha} \log \frac{e^{V_\psi(x, y_{<t}^{(i)})}}{\sum_{j=0}^{K-1} e^{V_\psi(x, y_{<t}^{(j)})}} \right]. \quad (7)$$

221 Intuitively, this objective trains V_ψ to assign higher values to those $y_{<t}$ that are likely to extend into
 222 high-reward outcomes. Considering the definition of $V(x, y_{<t})$ and the tower property, our objective
 223 is equivalent to minimizing a weighted group cross-entropy:
 224

$$225 \min_{V_\psi} \mathbb{E}_{\substack{x \sim \mathcal{D}, t \sim \mathcal{U}\{1, \dots, T\}, \\ y_{<t}^{(0:K)} \sim \pi_{\theta_{\text{old}}}(\cdot|x)}} \left[\left(\sum_{j=0}^{K-1} e^{V(x, y_{<t}^{(j)})} \right) \cdot \sum_{i=0}^{K-1} \frac{e^{V(x, y_{<t}^{(i)})}}{\sum_{j=0}^{K-1} e^{V(x, y_{<t}^{(j)})}} \log \frac{e^{V_\psi(x, y_{<t}^{(i)})}}{\sum_{j=0}^{K-1} e^{V_\psi(x, y_{<t}^{(j)})}} \right].$$

229 This equivalence reveals that the objective Eq. 7 frames the learning problem as one of distributional
 230 matching between two distributions defined over the group: (i) a target distribution derived from the
 231 normalized, exponential true values V , and (ii) a model distribution derived from the normalized,
 232 exponential estimated values of our learnable function V_ψ . From this observation, the following
 233 theorem establishes that the optimal solution to Eq. 7 recovers the true value function up to a constant
 234 offset.

235 **Theorem 3.2** (Consistency up to constant shift). *Assume unlimited model capacity and data. For any
 236 $K > 1$ and $t \in \{1, \dots, T\}$, the minimizer V_{ψ^*} of Eq. 7 recovers the soft value function V of Eq. 4
 237 up to an additive, $y_{<t}$ -independent offset $C_t(x)$:*

$$238 \quad V_{\psi^*}(x, y_{<t}) = V(x, y_{<t}) + \log C_t(x),$$

239 equivalently $e^{V_{\psi^*}(x, y_{<t})} = C_t(x) e^{V(x, y_{<t})}$.

241 This additive offset does not affect the resulting optimal policy, as in the following corollary.

243 **Corollary 3.3** (Policy invariance to additive shifts). *For any positive scalar $C_t(x)$, let $V'(x, y_{<t}) =$
 244 $V(x, y_{<t}) + \log C_t(x)$. The optimal policy for V' remains the same as the optimal policy for V .*

245 Taken together, these results guarantee that the optimal policy derived from our learned value function
 246 V_{ψ^*} is identical to the one obtained from the true value function V . The proofs are provided in
 247 Appendix A.

249 3.3 GROUP-NORMALIZED IMPLICIT VALUE OPTIMIZATION

251 The key step toward our final, a critic-free objective, is to eliminate the explicit value function V_ψ
 252 entirely. Instead of training a separate network, we leverage the policy-value link from Theorem 3.2
 253 by substituting the value term in Eq. 7 with its implicit, policy-defined equivalent.

254 From Theorem 3.2, we know that the optimal value function V_{ψ^*} is equivalent to the true value V
 255 up to a constant ($e^{V_{\psi^*}} = C_t e^V$). We can combine this with the policy-value link from Eq. 3, which
 256 states $\pi_{\theta^*}(y_{<t}|x) \propto \pi_{\theta_{\text{old}}}(y_{<t}|x) e^{V(x, y_{<t})}$. Together, these relationships allow us to express the
 257 exponential value function directly in terms of the policy:
 258

$$259 \quad e^{V_{\psi^*}(x, y_{<t})} = C_t(x) Z(x) \frac{\pi_{\theta^*}(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)}. \quad (8)$$

261 We now substitute this policy-defined expression for the exponential value function back into the
 262 model’s predicted distribution within our objective (Eq. 7). The model’s distribution is a softmax over
 263 the group of K sampled values. By replacing each $\exp(V_\psi)$ term with its policy-defined equivalent
 264 from Eq. 8, the shared multiplicative terms $C_t(x)$ and $Z(x)$ cancel out from the numerator and
 265 denominator of the softmax operation:

$$267 \quad \frac{\exp(V_\psi(y_{<t}^{(i)}))}{\sum_j \exp(V_\psi(y_{<t}^{(j)}))} \longrightarrow \frac{\cancel{C_t(x) Z(x)} \frac{\pi_{\theta}(y_{<t}^{(i)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(i)}|x)}}{\sum_{j=0}^{K-1} \cancel{C_t(x) Z(x)} \frac{\pi_{\theta}(y_{<t}^{(j)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(j)}|x)}} = \frac{\frac{\pi_{\theta}(y_{<t}^{(i)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(i)}|x)}}{\sum_{j=0}^{K-1} \frac{\pi_{\theta}(y_{<t}^{(j)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(j)}|x)}}.$$

270 This leads to our final, critic-free objective:

$$272 \quad \mathcal{L}_{\text{GN-IVO}}(\theta) = \mathbb{E}_{\substack{x \sim \mathcal{D}, t \sim \mathcal{U}\{1, \dots, T\}, \\ y^{(0:K)} \sim \pi_{\theta_{\text{old}}}(\cdot|x)}} \left[-\sum_{i=0}^{K-1} e^{R(x, y^{(i)})/\alpha} \left(\log \frac{\pi_{\theta}(y_{<t}^{(i)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(i)}|x)} - \log \sum_{j=0}^{K-1} \frac{\pi_{\theta}(y_{<t}^{(j)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(j)}|x)} \right) \right]. \quad (9)$$

276 In our practical implementation, we replace $e^{R(x, y^{(i)})/\alpha}$ with the group-normalized weight
 277 $\frac{e^{R(x, y^{(i)})/\alpha}}{\sum_{j=0}^{K-1} e^{R(x, y^{(j)})/\alpha}}$ for stability.
 278

280 **Training procedure** The training procedure for our algorithm follows the iterative scheme detailed
 281 in Algorithm 1. Each training iteration consists of data generation, reward evaluation, and policy
 282 optimization. First, we sample a batch of queries x from the dataset \mathcal{D} . For each query, we generate a
 283 set of K distinct responses by sampling from $\pi_{\theta_{\text{old}}}$, which is a frozen copy of the policy from the
 284 previous iteration. Next, each of these K full-sequence responses is evaluated by the reward function
 285 $R(x, y)$ to obtain a scalar reward. Finally, these queries, responses, and rewards are used to perform a
 286 gradient update on the policy parameter θ by minimizing Eq. 9. After the policy π_{θ} has been updated,
 287 $\pi_{\theta_{\text{old}}}$ is updated to match π_{θ} for data generation in the next iteration.
 288

Algorithm 1 Group-Normalized Implicit Value Optimization

290 **Input:** Reward function R , learning rate η , the policy π_{θ} and set $\pi_{\theta_{\text{old}}} \leftarrow \pi_{\theta}$
 291 **for** iterations **do**
 292 Sample a query $x \sim \mathcal{D}$ and generate K responses $y^{(0:K)} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$
 293 Evaluate reward $R(x, y^{(i)})$ for all $i \in \{0, \dots, K-1\}$
 294 Update θ by optimizing the following loss in Eq. 9, $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}_{\text{GN-IVO}}(\theta)$
 295 $\pi_{\theta_{\text{old}}} \leftarrow \pi_{\theta}$
 296 **end for**
 297

4 EXPERIMENTS

299 We evaluate our method on a diverse range of reasoning and text generation tasks. These evalua-
 300 tions include mathematical problem-solving with reasoning steps as well as open-ended language
 301 generation. The text generation tasks encompass creating helpful assistant responses, summarizing
 302 Reddit posts, and generating prompts for text-to-image models. Furthermore, we conduct a sensitivity
 303 analysis on hyperparameters.
 304

4.1 MATHEMATICAL REASONING

306 This task assesses a model’s ability to solve math problems that require multi-step reasoning, including
 307 generating intermediate derivations and the final answer.
 308

309 **Experiment setup** For RL training on mathematical reasoning, we use the MATH
 310 dataset (Hendrycks et al., 2021) and employ Qwen2.5-Math-7B (Yang et al., 2024a) and Llama-
 311 3.1-8B-Instruct (Meta-AI, 2024) as backbone models. Input prompts include an instruction and
 312 few-shot examples to elicit step-by-step derivations, where each step is explicitly marked with a label
 313 (e.g. step1, step2). In this setting, we define y_t as a single reasoning step; thus, y_t is a sequence of
 314 tokens that forms a complete step, and T is the total number of reasoning steps in the solution. The
 315 reward function is composed of two components, one for the correctness of the final answer and the
 316 other for adherence to the specified format. We evaluate performance on widely used mathematical
 317 reasoning benchmarks, including AMC 2023 (of Problem Solving, b), Minerva Math (Lewkowycz
 318 et al., 2022), Olympiad-Bench (He et al., 2024), and AIME 2024/2025 (of Problem Solving, a). These
 319 benchmarks cover a wide range of difficulties and reasoning depths, from competition-style problems
 320 to graduate-level questions. Further details are provided in the Appendix C.
 321

322 **Metrics and decoding** We report performance using the Pass@1 and Pass@3 metric. For Pass@1,
 323 we generate responses via greedy decoding. For Pass@3, we sample three completions for each
 324 problem using a temperature of 0.7. For all evaluations, we utilize the `math-verify` library to perform
 325 answer extraction and equivalence checking.
 326

324
 325
 326
 Table 1: Comparison of our method against baselines on the math reasoning task. The *Pass@3* ($P@3$)
 metric is calculated over three trials per query. **Bold** and underline indicate the best and second-best
 results, respectively.

Method	AMC2023		Minerva Math		Olympiad-Bench		AIME2024		AIME2025	
	P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3
<i>Llama-3.1-8B-Instruct</i>	27.5	37.5	<u>25.7</u>	32.7	15.6	24.2	3.3	10.0	<u>0.0</u>	0.0
SFT-winning	27.5	35.0	24.2	<u>35.6</u>	16.0	<u>27.1</u>	<u>6.6</u>	6.6	<u>0.0</u>	0.0
Online DPO	22.5	33.1	25.3	30.5	15.1	26.2	3.3	<u>13.3</u>	0.0	0.0
PPO	25.0	35.0	21.7	34.9	15.7	26.2	3.3	16.6	<u>3.3</u>	0.0
DRO	22.5	35.0	23.1	33.8	15.5	25.6	3.3	0.0	0.0	6.6
OREO	27.5	32.5	<u>25.7</u>	35.3	15.7	26.8	3.3	6.6	<u>0.0</u>	6.6
RLOO	<u>35.0</u>	<u>40.0</u>	26.1	34.1	<u>17.9</u>	26.1	<u>6.6</u>	16.6	0.0	0.0
GRPO	<u>35.0</u>	37.5	25.3	35.3	18.8	25.3	<u>6.6</u>	16.6	<u>3.3</u>	0.0
Ours	42.5	45.0	26.1	36.0	17.3	27.8	10.0	16.6	<u>3.3</u>	<u>3.3</u>
<i>Qwen2.5-Math-7B</i>	52.5	70.0	27.0	36.0	37.7	46.2	23.0	26.6	6.6	13.3
SFT-winning	57.5	62.5	30.5	36.0	38.8	49.0	23.3	26.6	13.3	13.3
Online DPO	57.5	70.0	27.2	37.1	36.2	<u>49.2</u>	23.3	30.0	<u>10.0</u>	13.3
PPO	47.5	67.5	28.3	37.8	38.6	49.0	23.3	30.0	<u>10.0</u>	13.3
DRO	55.0	67.5	<u>31.2</u>	37.1	37.7	48.7	23.3	33.3	<u>10.0</u>	13.3
OREO	55.0	70.0	31.6	38.6	38.5	49.1	16.6	30.0	<u>10.0</u>	13.3
RLOO	57.5	<u>72.5</u>	30.1	<u>40.4</u>	38.8	48.8	23.3	<u>36.6</u>	13.3	<u>16.6</u>
GRPO	60.0	70.0	29.7	37.8	<u>39.2</u>	49.4	<u>26.6</u>	33.3	6.6	13.3
Ours	62.5	75.0	31.6	41.9	39.8	49.0	30.0	40.0	<u>13.3</u>	23.3

343 **Baselines** We compare our proposed method against the following baseline algorithms:

344

- 345 • **SFT-winning** (Yuan et al., 2023) generates two responses for each query and performs supervised
 346 fine-tuning on the response with the higher reward.
- 347 • **Online DPO** (Qi et al., 2024) extends DPO to an online setting. In each iteration, it generates two
 348 responses for each query, designates them as the winner and the loser based on their rewards, and
 349 uses this pair to compute the DPO loss in the policy update.
- 350 • **PPO** (Schulman et al., 2017; Ouyang et al., 2022) generates a single completion for each query
 351 and performs clipped policy-gradient updates using advantages computed by a critic network.
- 352 • **DRO** (Richemond et al., 2024) is a soft Q-learning algorithm adapted to the bandit setting. It
 353 generates a single completion for each query and trains a value network via the soft Bellman
 354 equation. Policy updates are then derived from this value network.
- 355 • **OREO** (Wang et al., 2025) is a sequential extension of DRO that uses soft Q-learning to train a
 356 step-level value network. This network then guides the generation process at each step.
- 357 • **RLOO** (Ahmadian et al., 2024) is a REINFORCE-style policy gradient method that employs a
 358 leave-one-out estimation for advantages. It generates K responses for each query, and the rewards
 359 from this group are used to compute the leave-one-out advantage.
- 360 • **GRPO** (Shao et al., 2024) is a PPO-style method that uses the group mean as a baseline for
 361 advantage estimation. It generates K responses for each query, and the mean and standard-deviation
 362 of their rewards are used to compute the advantages.

363 We use the same group size (K) for GRPO, RLOO and our method across all experiments.

364

365 **Experimental results** The results are summarized in Table 1. Overall, our algorithm delivers
 366 consistent gains over strong RL and preference-learning baselines, and these improvements transfer
 367 across both model backbones. The performance differences can be explained in terms of each method’s
 368 approach. Methods that rely on an explicit value network, such as PPO, DRO, and OREO, showed
 369 limited improvement. While OREO models the task sequentially, its performance is likely constrained
 370 by the difficulty of accurately estimating value functions for long-range reasoning tasks. Similarly,
 371 Online DPO and Rejection Sampling, which use only two samples per query, also struggled. This is
 372 because in complex reasoning where most responses are incorrect, a simple pairwise preference signal
 373 provides an insufficient learning signal. In contrast, the methods that showed the most significant
 374 improvements, RLOO, GRPO, and our algorithm, all use a larger group of samples, thereby providing
 375 a meaningful learning signal. Among them, our algorithm consistently shows the best performance
 376 on most of the tasks. This is because our algorithm combines the benefits of a group-based signal
 377 with the fine-grained credit assignment of a sequential model, a feature lacking in other bandit-based
 methods.

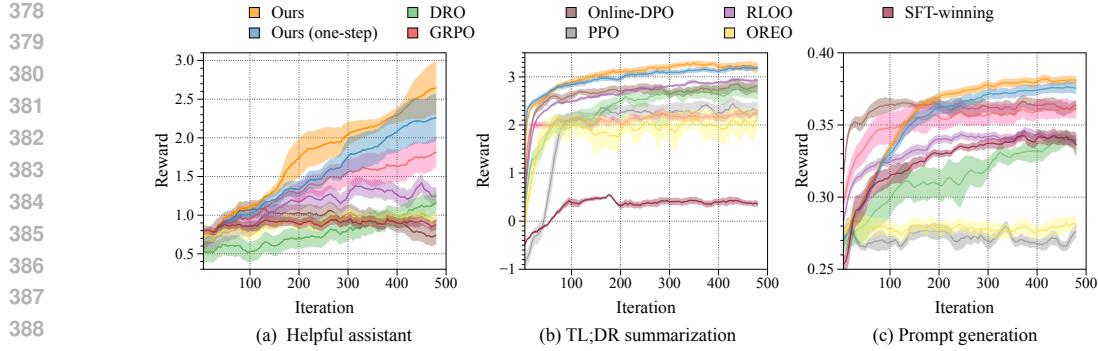


Figure 1: Training curves for our methods and baselines on the Llama-3.2-3B-Instruct model. The solid lines represent the mean reward, while the shaded regions indicate the standard deviation calculated over three random seeds.

4.2 TEXT GENERATION

We evaluate on three distinct text generation tasks: (1) Helpful assistant, (2) TL;DR summarization, and (3) Prompt generation for text-to-image models. For the Helpful assistant task, the model must generate responses that are both helpful and harmless, with rewards capturing human preferences for such responses. For the TL;DR summarization, the model is required to produce a concise summary of a given Reddit post. In the Prompt generation, the objective is to create a text prompt that guides a text-to-image model to generate an image in a specified artistic style.

Experiment setup Our experiments utilize two base models: Qwen2.5-1.5B-Instruct (Yang et al., 2024a) and Llama-3.2-3B-Instruct (Meta-AI., 2024). In these text generation tasks, the number of timesteps T corresponds to the sequence length. For the Helpful assistant task, we train on the Anthropic HH-RLHF (Bai et al., 2022) dataset and score responses using publicly available reward models from huggingface that assess helpfulness and harmlessness. For the TL;DR summarization, we use the TL;DR dataset (Stiennon et al., 2020) and evaluate the outputs using a publicly released model trained to assess summary quality. Finally, for the Prompt generation task, the model is given contextual instructions, such as an animal activity and a designated artistic style. The reward is calculated using the CLIP text-image similarity score between the generated prompt and the style reference images. Further details on dataset construction and the specific reward models used are provided in the Appendix C.

Metrics and decoding We report performance using two average reward metrics, Avg@1 and Avg@3. The Avg@1 score is calculated using greedy decoding. To compute Avg@3, we sample three completions for each query with a temperature of 0.7 and report the average of their rewards.

Experimental results We evaluate our primary sequential algorithm, where the number of steps T is the sequence length, alongside a one-step variant that treats the full generated sequence as a single step. Since OREO is a sequential extension of the DRO framework, we implement it with a token-level critic akin to that used in PPO. Training curves and final results are presented in Figure 1 and Table 2, respectively. [To further assess performance, we conduct an LLM evaluation, and results are summarized in Table 3](#). Generally, algorithms that use K samples per query outperform those using a single sample. While methods using multiple K samples are typically superior, even our one-step variant outperforms policy-gradient baselines like GRPO and RLOO. This result indicates that our group-normalized objective provides a more effective learning signal than standard policy-gradient estimation.

Our sequential algorithm generally outperforms its one-step variant across all tasks, except for summarization with Llama-3.2-3B-Instruct where performance is comparable due to the required short outputs. In contrast, we observe an opposite trend among DRO and OREO, where DRO typically outperforms OREO. We attribute this discrepancy to the challenge of training an accurate per-timestep value network for OREO, as an inaccurate critic can degrade policy performance. Because our method does not rely on an explicit auxiliary network, it successfully leverages the benefits of fine-grained credit assignment without being exposed to this pitfall.

4.3 ANALYSIS ON HYPERPARAMETERS

432

433 Table 2: Comparison of our methods against baselines on three text generation tasks. For each method,
 434 we run three random seeds and report the *best* scores on the test set across seeds. For *Avg@3*, we report
 435 the mean with std. over three repeated trials per query in the same model. *GM* denotes the geometric mean of
 436 *Avg@1* across the three tasks. **Bold** and underline indicate the best and the second-best accuracy for each task,
 437 respectively.

437

438 Method	439 Helpful assistant		440 TL;DR summarization		441 Prompt generation		442 GM
	443 Avg@1	444 Avg@3	445 Avg@1	446 Avg@3	447 Avg@1	448 Avg@3	
<i>Qwen2.5-1.5B-Instruct</i>							
SFT-winning	0.375	0.375 (0.001)	-0.710	-0.729 (0.008)	0.342	0.329 (0.004)	-0.450
Online DPO	1.271	1.260 (0.005)	0.998	0.660 (0.016)	<u>0.381</u> (0.000)	0.269 (0.005)	0.786
PPO	0.875	0.846 (0.003)	1.393	1.320 (0.015)	0.274	0.269 (0.005)	0.694
DRO	1.174	1.146 (0.007)	1.831	1.725 (0.015)	0.369	0.363 (0.004)	0.926
OREO	0.721	0.733 (0.002)	1.633	1.362 (0.033)	0.261	0.259 (0.003)	0.675
RLOO	1.120	1.043 (0.001)	2.466	2.423 (0.003)	0.319	0.311 (0.005)	1.004
GRPO	<u>1.594</u>	<u>1.506</u> (0.014)	1.151	1.086 (0.021)	0.367	0.367 (0.000)	0.876
Ours (one-step)	1.446	1.389 (0.005)	2.148	2.108 (0.024)	0.379	0.372 (0.001)	<u>1.056</u>
Ours	1.650	1.628 (0.009)	<u>2.418</u>	<u>2.359</u> (0.024)	0.383	0.382 (0.001)	1.152
<i>Llama-3.2-3B-Instruct</i>							
SFT-winning	0.819	0.819 (0.010)	0.603	0.897 (0.011)	0.354	0.354 (0.001)	0.640
Online DPO	1.064	0.974 (0.008)	2.907	2.884 (0.005)	<u>0.372</u>	<u>0.372</u> (0.000)	1.066
PPO	0.867	0.854 (0.013)	2.807	2.748 (0.016)	0.287	0.285 (0.002)	0.887
DRO	1.317	1.295 (0.002)	3.104	3.099 (0.008)	0.350	0.345 (0.002)	1.082
OREO	0.881	0.864 (0.008)	2.715	2.649 (0.018)	0.291	0.288 (0.005)	0.886
RLOO	1.528	1.262 (0.004)	3.181	3.116 (0.008)	0.349	0.345 (0.002)	1.193
GRPO	2.013	1.996 (0.013)	2.337	2.316 (0.001)	0.358	0.346 (0.005)	1.154
Ours (one-step)	<u>2.562</u>	<u>2.555</u> (0.002)	3.398	3.334 (0.007)	0.371	0.367 (0.001)	1.478
Ours	3.370	3.281 (0.005)	<u>3.347</u>	<u>3.330</u> (0.013)	0.384	0.382 (0.001)	1.630

455

456 Table 3: **LLM-based pairwise evaluation results.** We employed ChatGPT-4 to compare responses
 457 from our method against the baseline on the Qwen-2.5-1.5B-Instruct.

458

	459 Helpful assistant			460 TL;DR summarization			461 Prompt generation		
	462 win %	463 tie %	464 loss %	465 win %	466 tie %	467 loss %	468 win %	469 tie %	470 loss %
vs SFT-winning	99.8	0.0	0.1	99.8	0.0	0.1	100.0	0.0	0.0
vs Online DPO	64.9	20.6	14.4	75.6	0.1	24.1	45.1	20.3	34.4
vs PPO	90.5	0.3	9.1	98.8	0.0	1.1	52.5	0.3	47.1
vs DRO	58.5	30.4	10.9	50.3	3.8	45.8	17.1	16.8	66.0
vs OREO	75.8	0.6	23.5	51.6	4.9	43.4	51.4	0.0	48.5
vs RLOO	41.5	25.9	31.5	50.8	3.8	45.3	52.7	0.4	46.7
vs GRPO	56.3	1.6	33.0	55.8	10.0	34.1	59.3	20.0	20.6

475

476

The group size K We analyze the impact of the group size K , a key hyperparameter for both our method and GRPO, using the Qwen2.5-1.5B-Instruct model evaluated on the Anthropic HH-RLHF dataset. As illustrated in Figure 2.(a), the results show a clear trend: performance improves as K increases in both approaches. This behavior is expected, as our objective relies on the set of K samples to form an empirical target distribution. With a small K , this distribution is a noisy approximation of the true value landscape over all possible responses. As K grows, however, the target distribution more accurately reflects the relative ranking of responses within the group. This provides a more precise learning signal, allowing the model to learn a better value approximation. Notably, the performance gap between our method and GRPO is most pronounced when the group size is small, particularly at $K = 2$. This demonstrates that our algorithm is effective even in the small group size.

483

484

The temperature coefficient α We also analyze the impact of the temperature coefficient α , which controls the scale of rewards. The target distribution is formed with these scaled rewards, $\text{softmax}(R/\alpha)$, where a smaller α yields a sharper distribution. Conversely, a larger α makes the

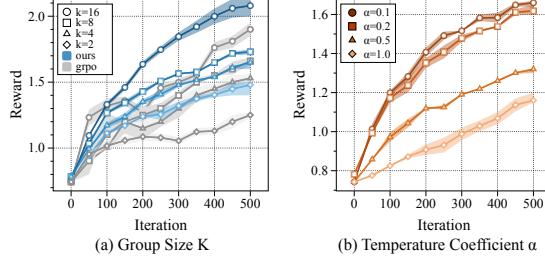


Figure 2: Analysis on hyperparameters

486 distribution softer and more uniform, reducing the distinction between responses with high and
 487 low rewards. We conduct experiments with $\alpha \in \{0.1, 0.2, 0.5, 1.0\}$, and the results are illustrated
 488 in Figure 2.(b). We observe that the model **reaches higher rewards** with lower temperature values
 489 ($\alpha = 0.1, 0.2$).

490 In contrast, for higher values such as $\alpha = 0.5$ and 1.0 , the models ultimately **yields lower rewards**.
 491 **This is because a more uniform target distribution provides a weaker and less discriminative signal**
 492 **over the group. As a result, the policy assigns comparable probability mass to both superior and**
 493 **mediocre responses, rather than differentiating them.**

495 5 CONCLUSION

496 We presented GN-IVO, a novel, critic-free RL algorithm for language models. Its core contribution is
 497 a group-normalized, distributional matching objective that learns step-level values implicitly from
 498 the policy itself, thereby bypassing the need for an explicit critic and its associated complexities.
 499 GN-IVO is theoretically grounded and demonstrates strong empirical performance on a diverse set
 500 of tasks. A promising future direction is to apply our framework to other generative models beyond
 501 language, such as for the fine-tuning of diffusion models.

504 505 THE USE OF LARGE LANGUAGE MODELS (LLMs)

506 LLMs were used to improve the clarity and readability of the text. Furthermore, they were utilized
 507 in the software development process for debugging code and identifying programming errors. All
 508 content generated by LLMs was critically reviewed and edited by the authors.

510 511 REFERENCES

512 Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin,
 513 Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for
 514 learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the
 515 Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, 2024.

516 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,
 517 Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion,
 518 Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan
 519 Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei,
 520 Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan.
 521 Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*,
 522 2022.

523 Yunseon Choi, Sangmin Bae, Seonghyun Ban, Minchan Jeong, Chuheng Zhang, Lei Song, Li Zhao,
 524 Jiang Bian, and Kee-Eung Kim. Hard prompts made interpretable: Sparse entropy regularization
 525 for prompt tuning with RL. In *Proceedings of the 62nd Annual Meeting of the Association for
 526 Computational Linguistics (Volume 1: Long Papers)*, ACL, 2024.

527 DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.
 528 *CoRR*, 2025.

530 Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: model
 531 alignment as prospect theoretic optimization. *CoRR*, 2024.

532 Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with
 533 deep energy-based policies. In *Proceedings of the 34th International Conference on Machine
 534 Learning, ICML*, 2017.

536 Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu,
 537 Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiad-
 538 bench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal
 539 scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computa-
 tional Linguistics (Volume 1: Long Papers)*, ACL, 2024.

540 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,
 541 and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In
 542 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks,*
 543 *NeurIPS Datasets and Benchmarks Track*, 2021.

544

545 Zijing Hu, Fengda Zhang, Long Chen, Kun Kuang, Jiahui Li, Kaifeng Gao, Jun Xiao, Xin Wang, and
 546 Wenwu Zhu. Towards better alignment: Training diffusion models with reinforcement learning
 547 against sparse rewards. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition,*
 548 *CVPR*. Computer Vision Foundation / IEEE, 2025.

549

550 Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron
 551 Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In
 552 *Forty-second International Conference on Machine Learning*, 2025.

553

554 Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal,
 555 Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Salman H. Khan, and Fahad Shahbaz Khan.
 556 LLM post-training: A deep dive into reasoning large language models. *CoRR*, 2025.

557

558 Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V.
 559 Ramasesh, Ambrose Sloane, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam
 560 Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language
 561 models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural*
 562 *Information Processing Systems, NeurIPS*, 2022.

563

564 Guanlin Liu, Kaixuan Ji, Renjie Zheng, Zheng Wu, Chen Dun, Quanquan Gu, and Lin Yan. Enhancing
 565 multi-step reasoning abilities of language models through direct Q-function optimization. *CoRR*,
 566 2024.

567

568 Joo Bon Maeng, Seongmin Lee, Seokin Seo, and Kee-Eung Kim. Goal-conditioned DPO: Priori-
 569 tizing safety in misaligned instructions. In *Proceedings of the 2025 Conference of the Nations*
 570 *of the Americas Chapter of the Association for Computational Linguistics: Human Language*
 571 *Technologies (Volume 1: Long Papers)*, NAACL, 2025.

572

573 Meta-AI. The llama 3 herd of models. *CoRR*, 2024.

574

575 Art of Problem Solving. Aime problems and solutions, a. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

576

577 Art of Problem Solving. Amc problems and solutions, b. URL https://artofproblemsolving.com/wiki/index.php?title=AMC_Problems_and_Solutions.

578

579 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong
 580 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton,
 581 Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and
 582 Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in*
 583 *Neural Information Processing Systems 35: Annual Conference on Neural Information Processing*
 584 *Systems, NeurIPS*, 2022.

585

586 Gaurav Pandey, Yatin Nandwani, Tahira Naseem, Mayank Mishra, Guangxuan Xu, Dinesh Raghu,
 587 Sachindra Joshi, Asim Munawar, and Ramón Fernandez Astudillo. BRAIn: Bayesian reward-
 588 conditioned amortized inference for natural language generation from feedback. In *Forty-first*
 589 *International Conference on Machine Learning*, 2024.

590

591 Biqing Qi, Pengfei Li, Fangyuan Li, Junqi Gao, Kaiyan Zhang, and Bowen Zhou. Online dpo: Online
 592 direct preference optimization with fast-slow chasing. *CoRR*, 2024.

593

594 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea
 595 Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In
 596 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information*
 597 *Processing Systems, NeurIPS*, 2023.

594 Pierre Harvey Richemond, Yunhao Tang, Daniel Guo, Daniele Calandriello, Mohammad Gheshlaghi
 595 Azar, Rafael Rafailov, Bernardo Ávila Pires, Eugene Tarassov, Lucas Spangher, Will Ellsworth,
 596 Aliaksei Severyn, Jonathan Mallinson, Lior Shani, Gil Shamir, Rishabh Joshi, Tianqi Liu, Rémi
 597 Munos, and Bilal Piot. Offline regularised reinforcement learning for large language models
 598 alignment. *CoRR*, 2024.

599 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 600 optimization algorithms. *CoRR*, 2017.

602 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 603 Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathemati-
 604 cal reasoning in open language models. *CoRR*, 2024.

605 Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec
 606 Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In
 607 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information
 608 Processing Systems, NeurIPS*, 2020.

609 Chaojie Wang, Yanchen Deng, Zhiyi Lv, Zeng Liang, Jujie He, Shuicheng Yan, and Bo An. Q*:
 610 Improving multi-step reasoning for llms with deliberative planning. *CoRR*, 2024.

612 Huajie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. Of-
 613 line reinforcement learning for LLM multi-step reasoning. In *Findings of the Association for
 614 Computational Linguistics: ACL*, 2025.

616 Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein.
 617 Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In
 618 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information
 619 Processing Systems, NeurIPS*, 2023.

620 An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu,
 621 Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu,
 622 Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert
 623 model via self-improvement. *CoRR*, 2024a.

624 Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-
 625 in-context: Multi-objective alignment of foundation models with dynamic preference adjustment.
 626 In *Forty-first International Conference on Machine Learning, ICML*, 2024b.

628 Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Boji Shan, Zeyuan Liu, Jia Deng,
 629 Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu,
 630 and Maosong Sun. Advancing LLM reasoning generalists with preference trees. In *The Thirteenth
 631 International Conference on Learning Representations, ICLR*, 2025.

632 Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling
 633 relationship on learning mathematical reasoning with large language models. *CoRR*, 2023.

635 Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman.
 636 Quiet-STar: Language models can teach themselves to think before speaking. In *First Conference
 637 on Language Modeling, COLM*, 2024.

638 Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level
 639 direct preference optimization. In *Forty-first International Conference on Machine Learning,
 640 ICML*, 2024.

641
 642
 643
 644
 645
 646
 647

648 A PROOFS
649

650 *Proof of Theorem 3.1.* We consider the process $x \mapsto y = (y_0, \dots, y_{T-1})$ with reward $R(x, y)$,
651 where the parameterized policy $\pi_\theta(y|x)$ is autoregressive model from the $\pi_\theta(y_t|y_{<t}, x)$ ($t \in$
652 $\{1, \dots, T\}$). We claim that the optimal parameterized policy $\pi_{\theta^*}(y|x)$ of this process (Eq. 2) deter-
653 mines the autoregressive policy $\pi_{\theta^*}(y_t|y_{<t}, x)$ as follows.

$$\begin{aligned}
 656 \quad \pi_{\theta^*}(y_{<t}|x) &= \sum_y \pi_{\theta^*}(y_{<t}|y, x) \pi_{\theta^*}(y|x) \\
 657 \\
 658 &= \sum_y \pi_{\theta^*}(y_{<t}|y, x) \pi_{\theta_{\text{old}}}(y|x) \frac{e^{R(x,y)/\alpha}}{Z(x)} \\
 659 \\
 660 &= \sum_y \pi_{\theta_{\text{old}}}(y_{<t}|y, x) \pi_{\theta_{\text{old}}}(y|x) \frac{e^{R(x,y)/\alpha}}{Z(x)} \quad (\pi_{\{\theta_{\text{old}}, \theta^*\}}(y_{<t}|y, x) = \delta(y_{<t}, y[:t])) \\
 661 \\
 662 &= \sum_y \pi_{\theta_{\text{old}}}(y_{<t}|x) \pi_{\theta_{\text{old}}}(y|y_{<t}, x) \frac{e^{R(x,y)/\alpha}}{Z(x)} \\
 663 \\
 664 &= \frac{\pi_{\theta_{\text{old}}}(y_{<t}|x)}{Z(x)} \underbrace{\sum_y \pi_{\theta_{\text{old}}}(y|y_{<t}, x) e^{R(x,y)/\alpha}}_{=e^{V(x,y_{<t})}} = \frac{\pi_{\theta_{\text{old}}}(y_{<t}|x)}{Z(x)} e^{V(x,y_{<t})}.
 \end{aligned} \tag{Eq. 2}$$

665 Therefore, as Theorem 3.1 suggests, the optimal autoregressive policy $\pi_{\theta^*}(y_{<t}|x)$ is a reweighted
666 policy by an exponential soft value function $V(x, y_{<t})$.
667

□

668 *Proof of Theorem 3.2.* The proof begins with the objective function defined in Eq. 7:
669

$$\min_{V_\psi} \mathbb{E}_{\substack{x \sim \mathcal{D}, t \sim \mathcal{U}\{1, \dots, T\}, \\ y^{(0:K)}_{<t} \sim \pi_{\theta_{\text{old}}}^{\text{joint}}(\cdot|x)}} \left[- \sum_{i=0}^{K-1} e^{R(x,y^{(i)})/\alpha} \log \frac{e^{V_\psi(x,y^{(i)}_{<t})}}{\sum_{j=0}^{K-1} e^{V_\psi(x,y^{(j)}_{<t})}} \right].$$

670 By applying the law of total expectation and decomposing the joint probability $\pi_{\theta_{\text{old}}}(y_{<t}, y|x)$ into
671 $\pi_{\theta_{\text{old}}}(y_{<t}|x) \pi_{\theta_{\text{old}}}(y|y_{<t}, x)$, we can rewrite the expectation as:
672

$$\min_{V_\psi} \mathbb{E}_{\substack{x \sim \mathcal{D}, t \sim \mathcal{U}\{1, \dots, T\}, \\ y^{(0:K)}_{<t} \sim \pi_{\theta_{\text{old}}}(\cdot|x)}} \left[- \sum_{i=0}^{K-1} \mathbb{E}_{\pi_{\theta_{\text{old}}}(y^{(i)}|y^{(i)}_{<t}, x)} \left[e^{R(x,y^{(i)})/\alpha} \right] \log \frac{e^{V_\psi(x,y^{(i)}_{<t})}}{\sum_{j=0}^{K-1} e^{V_\psi(x,y^{(j)}_{<t})}} \right].$$

673 By definition, the soft value function satisfies $e^{V(x,y^{(i)}_{<t})} = \mathbb{E}_{\pi_{\theta_{\text{old}}}(y^{(i)}|y^{(i)}_{<t}, x)} \left[e^{R(x,y^{(i)})/\alpha} \right]$. Substitut-
674 ing this into the objective yields:
675

$$\min_{V_\psi} \mathbb{E}_{\substack{x \sim \mathcal{D}, t \sim \mathcal{U}\{1, \dots, T\}, \\ y^{(0:K)}_{<t} \sim \pi_{\theta_{\text{old}}}(\cdot|x)}} \left[- \sum_{i=0}^{K-1} e^{V(x,y^{(i)}_{<t})} \log \frac{e^{V_\psi(x,y^{(i)}_{<t})}}{\sum_{j=0}^{K-1} e^{V_\psi(x,y^{(j)}_{<t})}} \right].$$

676 To reveal the underlying structure, we can rewrite the expression as a weighted cross-entropy, which
677 is equivalent to the objective in Eq. 3.2:
678

$$\min_{V_\psi} \mathbb{E}_{\substack{x \sim \mathcal{D}, t \sim \mathcal{U}\{1, \dots, T\}, \\ y^{(0:K)}_{<t} \sim \pi_{\theta_{\text{old}}}(\cdot|x)}} \left[- \left(\sum_{j=0}^{K-1} e^{V(x,y^{(j)}_{<t})} \right) \sum_{i=0}^{K-1} \frac{e^{V(x,y^{(i)}_{<t})}}{\sum_{j=0}^{K-1} e^{V(x,y^{(j)}_{<t})}} \log \frac{e^{V_\psi(x,y^{(i)}_{<t})}}{\sum_{j=0}^{K-1} e^{V_\psi(x,y^{(j)}_{<t})}} \right].$$

679 For clarity, let us define two discrete probability distributions over the indices $i = 0, \dots, K-1$:
680

$$p_i(x, y^{(0:K)}_{<t}) := \frac{e^{V(x,y^{(i)}_{<t})}}{\sum_{j=0}^{K-1} e^{V(x,y^{(j)}_{<t})}} \quad \text{and} \quad q_{\psi,i}(x, y^{(0:K)}_{<t}) := \frac{e^{V_\psi(x,y^{(i)}_{<t})}}{\sum_{j=0}^{K-1} e^{V_\psi(x,y^{(j)}_{<t})}}.$$

Letting $A(x, y_{<t}^{(0:K)}) := \sum_{j=0}^{K-1} e^{V(x, y_{<t}^{(j)})}$, we can express the objective as:

$$\min_{\psi} \mathbb{E}_{t, \pi_{\theta_{\text{old}}}(y_{<t}^{(0:K)}|x)} \left[-A(x, y_{<t}^{(0:K)}) \sum_{i=0}^{K-1} p_i(x, y_{<t}^{(0:K)}) \log q_{\psi,i}(x, y_{<t}^{(0:K)}) \right].$$

The term inside the expectation is a positive scaling factor $A(\cdot)$ multiplied by the negative cross-entropy between distributions p and q . By Gibbs' inequality, this term is minimized when the two distributions are identical:

$$-A(\cdot) \sum_{i=0}^{K-1} p_i(\cdot) \log q_{\psi,i}(\cdot) \geq -A(\cdot) \sum_{i=0}^{K-1} p_i(\cdot) \log p_i(\cdot).$$

The minimum is achieved when $q_{\psi,i}(x, y_{<t}^{(0:K)}) = p_i(x, y_{<t}^{(0:K)})$ for all $i = 0, \dots, K-1$. Assuming sufficient model capacity and data, the optimal parameters ψ^* will satisfy this equality:

$$\frac{e^{V_{\psi^*}(x, y_{<t}^{(i)})}}{\sum_{j=0}^{K-1} e^{V_{\psi^*}(x, y_{<t}^{(j)})}} = \frac{e^{V(x, y_{<t}^{(i)})}}{\sum_{j=0}^{K-1} e^{V(x, y_{<t}^{(j)})}},$$

for any set of samples $\{y_{<t}^{(i)}\}_{i=0}^{K-1}$ from the support of $\pi_{\theta_{\text{old}}}(\cdot|x)$. This implies that for any pair of indices $i, k \in \{0, \dots, K-1\}$, the ratio of the exponentiated value functions is constant:

$$\frac{e^{V_{\psi^*}(x, y_{<t}^{(i)})}}{e^{V(x, y_{<t}^{(i)})}} = \frac{e^{V_{\psi^*}(x, y_{<t}^{(k)})}}{e^{V(x, y_{<t}^{(k)})}} = \frac{\sum_{j=0}^{K-1} e^{V_{\psi^*}(x, y_{<t}^{(j)})}}{\sum_{j=0}^{K-1} e^{V(x, y_{<t}^{(j)})}} =: \tilde{C}_t(x, y_{<t}^{(0:K)}).$$

The term \tilde{C}_t is constant for a given set of samples $\{y_{<t}^{(j)}\}_{j=0}^{K-1}$. We now show that this constant is independent of the particular choice of samples and can be written as $C_t(x)$. Consider any two partial sequences $y'_{<t}$ and $y''_{<t}$ in the support of $\pi_{\theta_{\text{old}}}(\cdot|x)$. We can construct a set of samples $\{y_{<t}^{(j)}\}_{j=0}^{K-1}$ that includes both, for instance by setting $y_{<t}^{(0)} = y'_{<t}$ and $y_{<t}^{(1)} = y''_{<t}$. Applying the result above with $i = 0$ and $k = 1$ gives:

$$\frac{e^{V_{\psi^*}(x, y'_{<t})}}{e^{V(x, y'_{<t})}} = \frac{e^{V_{\psi^*}(x, y''_{<t})}}{e^{V(x, y''_{<t})}}.$$

Since $y'_{<t}$ and $y''_{<t}$ are arbitrary sequences from the support, this ratio must be constant for any $y_{<t} \in \text{supp}(\pi_{\theta_{\text{old}}}(\cdot|x))$.

Therefore, there exists a function $C_t(x)$, which depends on t and x but not on $y_{<t}$, such that for all $y_{<t} \in \text{supp}(\pi_{\theta_{\text{old}}}(\cdot|x))$:

$$e^{V_{\psi^*}(x, y_{<t})} = C_t(x) \cdot e^{V(x, y_{<t})}.$$

□

Proof of Corollary 3.3. We consider the policy π' induced by the shifted value function V' , following the form in Eq. 3. We substitute the condition $e^{V'(x, y_{<t})} = C_t(x) \cdot e^{V(x, y_{<t})}$ into the policy definition:

$$\frac{\pi_{\theta_{\text{old}}}(y_{<t}|x) e^{V'(x, y_{<t})}}{\sum_{y_{<t}} \pi_{\theta_{\text{old}}}(y_{<t}|x) e^{V'(x, y_{<t})}} = \frac{C_t(x) \pi_{\theta_{\text{old}}}(y_{<t}|x) e^{V(x, y_{<t})}}{C_t(x) \sum_{y_{<t}} \pi_{\theta_{\text{old}}}(y_{<t}|x) e^{V(x, y_{<t})}} = \pi_{\theta^*}(y_{<t}|x).$$

Since the term $C_t(x)$ does not depend on the integration variable $y_{<t}$, it can be factored out of the integral in the denominator and cancels. Thus, the induced policy is identical to the original optimal policy. □

B COMPARISON WITH DRO, OREO

Comparison with DRO (Richemond et al., 2024) DRO treats the language generation as a bandit task, where the entire generation y is considered a one-step action. It minimizes the Soft Bellman error, which is the same with Eq. 6. In this setting, the expectation over future steps in the soft value

756 function collapses to the direct reward $R(x, y)/\alpha$. The MSE loss from Eq. 6 therefore simplifies to
 757 the following loss:
 758

$$759 \quad \mathcal{L}_{\text{DRO}}(\psi, \theta) = \left(\log Z_\psi(x) + \log \frac{\pi_\theta(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)} - R(x, y)/\alpha \right)^2. \quad (10)$$

761 The gradient of \mathcal{L}_{DRO} w.r.t. θ takes a form:
 762

$$763 \quad \nabla_\theta \mathcal{L}_{\text{DRO}}(\psi, \theta) = - (R(x, y) - \log Z_\psi(x)) \nabla_\theta \log \pi_\theta(y|x) - \frac{\alpha}{2} \nabla_\theta \left(\log \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)} \right)^2. \quad (11)$$

766 In this objective, $\log Z_\psi(x)$ acts as a baseline to reduce variance, and the second term is a regularization
 767 penalty. Thus, DRO’s training procedure involves optimizing two distinct objectives: Eq. 10 for
 768 the value function and Eq. 11 for the policy.
 769

770 **Comparison with OREO (Wang et al., 2025)** OREO extends DRO to a sequential decision making
 771 setting, which allows step-level credit assignment. It trains its step-wise value function, $V_\psi(x, y_{<t})$,
 772 by minimizing a Soft Bellman error, which takes the form of the MSE loss:
 773

$$774 \quad \mathcal{L}_{\text{OREO}}(\psi, \theta) = \left(V_\psi(x, y_{<t}) + \log \frac{\pi_\theta(y|y_{<t}, x)}{\pi_{\theta_{\text{old}}}(y|y_{<t}, x)} - R(x, y)/\alpha \right)^2. \quad (12)$$

776 Similar to DRO, the gradient of $\mathcal{L}_{\text{OREO}}$ w.r.t. θ takes the form:
 777

$$778 \quad \nabla_\theta \mathcal{L}_{\text{OREO}}(\psi, \theta) = - (R(x, y) - V_\psi(x, y_{<t})) \nabla_\theta \log \pi_\theta(y|y_{<t}, x) - \frac{\alpha}{2} \nabla_\theta \left(\log \frac{\pi_\theta(y|y_{<t}, x)}{\pi_{\theta_{\text{old}}}(y|y_{<t}, x)} \right)^2. \quad (13)$$

780 In practice, the policy gradient is calculated only with respect to the action taken at the current step
 781 $\pi_\theta(y_t|y_{<t}, x)$, and `.detach()` is applied to the future-step decisions, $\sum_{i=t+1}^T \log \pi_\theta(y_i|y_{<i}, x)$.
 782

783 In contrast, our approach does not update the policy based on a single action at each timestep t .
 784 Instead, it learns by considering the entire prefix $y_{<t}$, which allows the model to implicitly learn the
 785 value function $V(x, y_{<t})$ and directly assess the quality of partial sequences.
 786

787 **Online vs. Offline** DRO and OREO are off-policy algorithms that primarily target offline settings
 788 (learning from a fixed dataset). Our algorithm is also off-policy, as the behavior policy ($\pi_{\theta_{\text{old}}}$) can
 789 differ from the target policy (π_θ). However, unlike those methods, it is particularly well-suited for an
 790 online setting, as suitable offline datasets with the required K samples per query are rare.
 791

792 C TASKS DETAILS

794 C.1 MATHEMATICAL REASONING

796 To ensure all models generate structured, step-by-step solutions for the mathematical reasoning task,
 797 we utilize a instruction template. This template, shown in Figure 3, guides the model to break down
 798 its reasoning, show intermediate steps, verify its solution, and provide a final answer in a specific
 799 format. We use two types of rewards: one based on the correctness of the final answer, and another
 800 based on proper formatting, specifically checking the presence of `\boxed{}` in the response.
 801

802 C.2 TEXT GENERATION

803 **Helpful assistant** We use the Anthropic HH-RLHF dataset (Bai et al., 2022) which contains human
 804 preference data on helpful and harmless AI assistants. For the reward models, we adopt open source
 805 reward models: one for helpfulness³ and another for harmlessness⁴ following the setting in (Yang
 806 et al., 2024b). The final reward is a simple average of the two model scores, with each component
 807 equally weighted at 0.5.
 808

809 ³Ray2333/gpt2-large-helpful-reward_model

⁴Ray2333/gpt2-large-harmless-reward_model

```

810 Solve the following math problem step by step.
811 - Write each reasoning step clearly, starting with labels step1, step2, step3, ...
812 - Show intermediate formulas and simplifications.
813 - Before giving the final answer, add a verification step:
814 - You may use Python code, an alternative formula, or a quick logical check to confirm correctness.
815 - If you detect a mistake, correct it.
816 - Conclude with the final answer inside LaTeX  $\square$ .
817 Format example:
818 step1: Restate the problem.
819 step2: Apply the relevant formulas.
820 step3: Compute intermediate results.
821 step4: Interpret the results.
822 step5: Verify the solution (with Python code, alternative calculation, or logical consistency check).
823 step6: State the final boxed answer.
824
825 

---

 Few shot examples 

---


826 Problem: {Problem}
827 Assistant:

```

Figure 3: Instruction template used in mathematical reasoning task.

TL;DR summarization For this task, we use the Reddit TL;DR summarization (Stiennon et al., 2020) dataset. Following the setup in (Yang et al., 2024b), we construct a reward function using two open source models: one that rewards conciseness⁵ and another that rewards faithfulness to the source text⁶. The final reward is calculated as a simple average of the two model scores, with each component equally weighted at 0.5.

Prompt generation for text-to-image models Inspired by prior works on controlling text-to-images models through prompt generation (Wen et al., 2023; Choi et al., 2024), we designed a dataset for this task. The dataset is constructed by combining 856 animal activity scene descriptions from (Hu et al., 2025) with 4 distinct artistic styles (Surrealism, East Asian Classics, and Impressionist). It results in a total of 2559 (853 \times 3) scene-style pairs. The dataset is split evenly into training and test sets, with 1,712 pairs in each. Each artistic style is represented by three reference images, as illustrated in Figure 4. The examples of animal activity scene descriptions are shown in Table 4. Given a scene description and a target style, the model’s task is to generate a text prompt that can guide text-to-image models to produce an image in the required style while preserving the original scene’s context. The reward function consists of two components: a style score and a context preservation score. The style score is calculated using a frozen CLIP model⁷. We compute the CLIP similarity between the generated prompt (via the text encoder) and each of the three reference images (via the image encoder), and then average the results. The context preservation score is the cosine similarity between the generated prompt and the original scene description. The final reward is a weighted sum of these two components with weights of 0.7 for the style score and 0.3 for the context preservation score. All queries are formatted using the instruction template shown in Figure 5.

D IMPLEMENTATION DETAILS

All methods are implemented using the `trl` library. For response generation during training, we apply a sampling temperature of 0.7, top-p of 1.0, and top-k of 0.0. We also incorporate a KL-penalty regularization term between the current policy and the initial pretrained model, controlled by a coefficient β .

The mathematical reasoning experiments were conducted on 8 NVIDIA H200 GPUs with a training query batch size of 4 per GPU. We trained all methods for 4 epochs. During training, the maximum

⁵Tristan/gpt2.reward.summarization

⁶CogComp/bart-faithful-summary-detector

⁷openai/clip-vit-large-patch14

864

865

866

Table 4: Example of animal activity scene description for prompt generation task.

Example of animal activity scene description		
a cat washing dishes	a dog washing dishes	a duck washing dishes
a monkey riding a bike	a horse riding a bike	a pig riding a bike
a spider playing chess	a rabbit playing chess	a dolphin playing chess
a deer reading a book	a zebra reading a book	a ant reading a book
a lion washing dishes	a cow washing dishes	a turtle washing dishes
a raccoon cooking dinner	a bird cooking dinner	a llama cooking dinner
a lizard drawing a picture	a kangaroo drawing a picture	a shark drawing a picture
a butterfly playing the piano	a gorilla playing the piano	a lizard playing the piano
a whale writing a letter	a bee writing a letter	a fox writing a letter
a mouse jogging in the park	a chicken jogging in the park	a bear jogging in the park

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

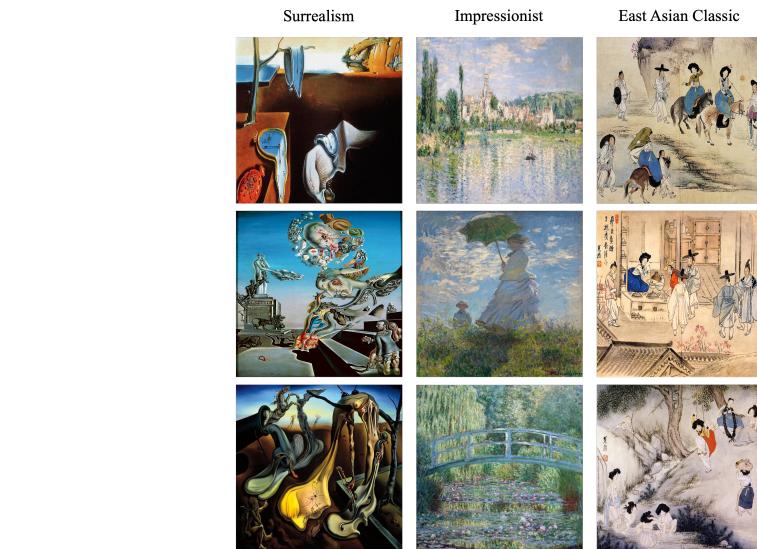


Figure 4: Reference images of 3 distinct artistic style for prompt generation task.

897

898

899

900

901

902

903

904

You are a prompt engineer for Stable Diffusion/SDXL.
Given Scene/Style, output one descriptive single-line prompt.

Scene: {Animal Activity Scene Description}
Scene: {Style}

Prompt:

905

906

907

new tokens was set to 1024, this was increased to 2048 for evaluation to accommodate the longer responses required by difficult benchmarks.

The text generation experiments were conducted on 4 NVIDIA H100 GPUs using a per-GPU query batch size of 8 and were trained for 500 iterations. The maximum generation length was set to 256 new tokens.

913

914

915

916

917

Model architecture All methods share the same policy network architecture, using either a Qwen or Llama model as the backbone. For critic-based algorithms, a separate value network was infeasible due to GPU memory limitations. We therefore utilized a ValueHead from the `tr1` library, which attaches head to the policy network. For the reasoning task, we fine-tune both the Qwen2.5-Math-7B and Llama-3.1-8B-Instruct models using a LoRA adapter.

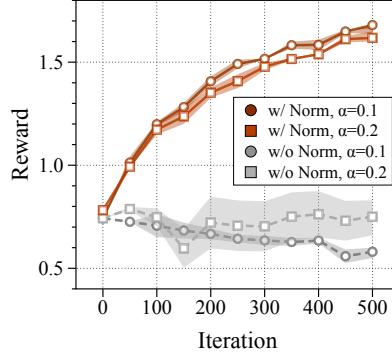
918 **SFT-winning** For each query in a batch, we generate two distinct responses ($K = 2$). Each response
 919 is then evaluated using a task-specific reward function, which yields a scalar score. The response with
 920 the higher score is selected as the winner. In the case of a tie, the first-generated response is chosen.
 921 Finally, the model’s policy is updated by maximizing the log-likelihood of generating these winning
 922 responses.

923 **Online DPO** For each query, we generate two completions ($K = 2$). These are then evaluated with
 924 a scalar reward function to form a preference pair, designating the response with the higher score as
 925 the winner and the other as the loser. In the event of a tie, the first-generated response is selected as
 926 the winner. Finally, the model’s policy is updated by applying the DPO loss to this constructed pair.

927 **DRO and OREO** While DRO and OREO are primarily designed for offline settings, we adapt
 928 them for online learning. We implement both DRO and OREO using the PPO trainer from the `tr1`
 929 library. The critic is implemented as a value head attached to the policy model. For training the value
 930 network, we use the corresponding loss functions defined in Eq. 10 for DRO and Eq. 12 for OREO.
 931 The policy is updated using the standard clipped surrogate objective from PPO, with advantages
 932 estimated by the learned value network. While we also experimented with implementing the policy
 933 losses directly from Eq. 11 and Eq. 13, we found that this approach led to unstable training and
 934 yielded poor performance.

935 **PPO, GRPO and RLOO** We implement the PPO, GRPO, and RLOO based on their respective
 936 trainers from the `tr1` library.

937 **GN-IVO (ours)** Our method, GN-IVO, is implemented using the base trainer from the `tr1` library
 938 and we use the same group size (K) as the GRPO and RLOO baselines for a fair comparison. As
 939 mentioned in Section 3.3, we use the normalized exponential reward rather than the exponential
 940 reward from Eq. 9; The impact of this normalization is analyzed in Figure 7.



945
 946 Figure 6: Investigating the impact of normalization on rewards from the Qwen2.5-1.5B-Instruct
 947 model. This comparison conducted α values of 0.1 and 0.2.

948 The detail of hyperparameters are provided in Table 5.

949
 950 **E QUALITATIVE RESULTS**

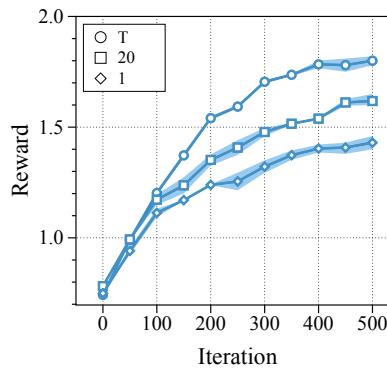


Figure 7: Investigating the impact of the number of sampled t from the Qwen2.5-1.5B-Instruct model. This comparison conducted values across $\{1, 20, T\}$.

Table 5: Hyperparameters used for algorithms across tasks.

Method	Hyperparameter	Reasoning (Llama3.1)	Reasoning (Qwen2.5)	Helpful assistant	TL;DR summarization	Prompt generation
SFT-winning	learning rate η	0.000001	0.000001	0.00001	0.00001	0.00001
	gradient clip norm	0.1	0.1	0.5	0.5	0.5
	KL coeff with init-model β	0.05	0.01	0.1	0.1	0.1
Online DPO	learning rate η	0.000001	0.000001	0.00001	0.00001	0.00001
	gradient clip norm	0.1	0.1	0.5	0.5	0.5
	temperature α	0.2	0.2	0.2	0.2	0.2
PPO	KL coeff with init-model β	0.05	0.01	0.1	0.1	0.1
	learning rate η	0.000001	0.000001	0.00001	0.00001	0.00001
	gradient clip norm	0.1	0.1	0.5	0.5	0.5
DRO	clipping range	0.2	0.2	0.2	0.2	0.2
	temperature α	0.2	0.2	0.2	0.2	0.2
	KL coeff with init-model β	0.05	0.01	0.1	0.1	0.1
OREO	learning rate η	0.000001	0.000001	0.00001	0.00001	0.00001
	gradient clip norm	0.1	0.1	0.5	0.5	0.5
	clipping range	0.2	0.2	0.2	0.2	0.2
RLOO	temperature α	0.2	0.2	0.2	0.2	0.2
	KL coeff with init-model β	0.05	0.01	0.1	0.1	0.1
	learning rate η	0.000001	0.000001	0.00005	0.00005	0.00005
GRPO	gradient clip norm	0.1	0.1	0.5	0.5	0.5
	group size K	4	4	4	4	4
	KL coeff with init-model β	0.05	0.01	0.2	0.2	0.2
Ours	learning rate η	0.000001	0.000001	0.00001	0.00001	0.00001
	gradient clip norm	0.1	0.1	0.5	0.5	0.5
	group size K	4	4	4	4	4
Ours	temperature α	0.2	0.2	0.2	0.2	0.2
	KL coeff with init-model β	0.05	0.01	0.1	0.1	0.1

1026
1027

E.1 MATHEMATICAL REASONING

1028
1029

Table 6: Example qualitative results generated by the Qwen2.5-Math-7B model on the AIME 2024.

1030
1031
1032
1033
1034
1035
1036

Question: Every morning Aya goes for a 9-kilometer-long walk and stops at a coffee shop afterwards. When she walks at a constant speed of s kilometers per hour, the walk takes her 4 hours, including t minutes spent in the coffee shop. When she walks $s + 2$ kilometers per hour, the walk takes her 2 hours and 24 minutes, including t minutes spent in the coffee shop. Suppose Aya walks at $s + \frac{1}{2}$ kilometers per hour. Find the number of minutes the walk takes her, including the t minutes spent in the coffee shop.

1037
1038

Method	Response
Ours	<p>step1: Restate the problem. Aya walks 9 kilometers at two different speeds, s and $s+2$, and stops for t minutes in a coffee shop. We need to find the time it takes her to walk 9 kilometers at a speed of $s + \frac{1}{2}$ kilometers per hour, including the t minutes spent in the coffee shop.</p> <p>step2: Set up the equations. When Aya walks at speed s km/h, the total time (including the coffee shop stop) is 4 hours. So the walking time is $4 - \frac{t}{60}$ hours. The distance is 9 km, so we have:</p> $9 = s \left(4 - \frac{t}{60} \right)$ <p>When Aya walks at speed $s + 2$ km/h, the total time (including the coffee shop stop) is 2 hours and 24 minutes, which is $2 + \frac{24}{60} = 2.4$ hours. So the walking time is $2.4 - \frac{t}{60}$ hours. The distance is 9 km, so we have:</p> $9 = (s + 2) \left(2.4 - \frac{t}{60} \right)$ <p>step3: Solve the system of equations. We can solve these equations to find s and t. First, let's express t in terms of s from the first equation:</p> $9 = s \left(4 - \frac{t}{60} \right) \implies 9 = 4s - \frac{st}{60} \implies 540 = 240s - st$ $\implies t = \frac{240s - 540}{s}$ <p>Next, substitute t into the second equation:</p> $9 = (s + 2) \left(2.4 - \frac{t}{60} \right)$ $9 = (s + 2) \left(2.4 - \frac{240s - 540}{60s} \right)$ $9 = (s + 2) \left(2.4 - \frac{4s - 9}{s} \right)$ $9 = (s + 2) \left(\frac{2.4s - 4s + 9}{s} \right)$ $9 = (s + 2) \left(\frac{9 - 1.6s}{s} \right)$

1078

1079

Continued on next page

1080	Method	Response
1081		
1082		
1083		$9 = \frac{(s+2)(9-1.6s)}{s}$
1084		$9s = (s+2)(9-1.6s)$
1085		$9s = 9s - 1.6s^2 + 18 - 3.2s$
1086		$0 = -1.6s^2 + 18 - 3.2s$
1087		$1.6s^2 + 3.2s - 18 = 0$
1088		$s^2 + 2s - 11.25 = 0$
1089		
1090		
1091		We can solve this quadratic equation using the quadratic formula $s = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, where $a = 1$, $b = 2$, and $c = -11.25$:
1092		
1093		
1094		$s = \frac{-2 \pm \sqrt{4 + 45}}{2} = \frac{-2 \pm 7}{2}$
1095		
1096		The positive solution is $s = \frac{5}{2} = 2.5$.
1097		Now we can find t :
1098		
1099		$t = \frac{240s - 540}{s} = \frac{240 \cdot 2.5 - 540}{2.5} = \frac{600 - 540}{2.5} = \frac{60}{2.5} = 24$
1100		
1101		
1102		step4: Find the time for the speed $s + \frac{1}{2}$.
1103		The speed is $s + \frac{1}{2} = 2.5 + 0.5 = 3$ km/h. The walking time is:
1104		
1105		$\text{walking time} = \frac{9}{3} = 3 \text{ hours}$
1106		
1107		Including the coffee shop stop, the total time is:
1108		
1109		$3 + \frac{24}{60} = 3 + 0.4 = 3.4 \text{ hours}$
1110		
1111		Converting 0.4 hours to minutes, we get $0.4 \times 60 = 24$ minutes. So the total time is 3 hours 24 minutes.
1112		
1113		Therefore, the number of minutes the walk takes her, including the t minutes spent in the coffee shop, is:
1114		
1115		$3 \times 60 + 24 = 180 + 24 = 204 \text{ minutes}$
1116		
1117		step5: State the final boxed answer.
1118		
1119		204
1120		
1121		
1122		
1123		
1124		
1125		
1126		
1127		
1128		
1129		
1130		
1131		
1132		
1133		

1134 Table 7: Example qualitative results generated by the Qwen2.5-Math-7B model on the AIME 2025.
1135

1136 **Question:** *An isosceles trapezoid has an inscribed circle tangent to each of its four sides. The*
 1137 *radius of the circle is 3, and the area of the trapezoid is 72. Let the parallel sides of the trapezoid*
 1138 *have lengths r and s , with $r \neq s$. Find $r^2 + s^2$.*

Method	Response
Ours	<p>1143 step1: Restate the problem. 1144 We have an isosceles trapezoid with an inscribed circle of radius 3. The area of 1145 the trapezoid is 72. We need to find $r^2 + s^2$ where r and s are the lengths of the 1146 parallel sides of the trapezoid. 1147 step2: Use the properties of the trapezoid and the inscribed circle. 1148 For a trapezoid to have an inscribed circle, the sum of the lengths of the parallel 1149 sides must equal the sum of the lengths of the non-parallel sides. Let the lengths 1150 of the parallel sides be r and s with $r > s$, and let the lengths of the non-parallel 1151 sides be a. Then we have: 1152 $r + s = 2a$ 1153 The height of the trapezoid is equal to the diameter of the inscribed circle, which 1154 is $2 \times 3 = 6$. 1155 The area of the trapezoid is given by: 1156 $\text{Area} = \frac{1}{2} \times (r + s) \times \text{height} = \frac{1}{2} \times (r + s) \times 6 = 3(r + s)$ 1157 We know the area is 72, so: 1158 $3(r + s) = 72$ 1159 $r + s = 24$ 1160 From the earlier equation $r + s = 2a$, we get: 1161 $2a = 24$ 1162 $a = 12$ 1163 step3: Use the Pythagorean theorem to find the lengths of the legs of the right 1164 triangles formed by dropping perpendiculars from the endpoints of the shorter 1165 base to the longer base. 1166 The trapezoid can be divided into a rectangle and two right triangles. The legs of 1167 the right triangles are the height of the trapezoid (6) and half the difference of the 1168 bases $\frac{r-s}{2}$. The hypotenuse of each right triangle is $a = 12$. So we have: 1169 $12^2 = 6^2 + \left(\frac{r-s}{2}\right)^2$ 1170 $144 = 36 + \left(\frac{r-s}{2}\right)^2$ 1171 $108 = \left(\frac{r-s}{2}\right)^2$ 1172 $\frac{r-s}{2} = \sqrt{108} = 6\sqrt{3}$ 1173 $r-s = 12\sqrt{3}$</p>

1184 Continued on next page
1185

Method	Response
1188	
1189	
1190	step4: Solve the system of equations $r + s = 24$ and $r - s = 12\sqrt{3}$.
1191	Adding the two equations:
1192	$(r + s) + (r - s) = 24 + 12\sqrt{3}$
1193	$2r = 24 + 12\sqrt{3}$
1194	$r = 12 + 6\sqrt{3}$
1195	
1196	
1197	Subtracting the two equations:
1198	$(r + s) - (r - s) = 24 - 12\sqrt{3}$
1199	$2s = 24 - 12\sqrt{3}$
1200	$s = 12 - 6\sqrt{3}$
1201	
1202	step5: Calculate $r^2 + s^2$.
1203	
1204	$r^2 = (12 + 6\sqrt{3})^2 = 144 + 144\sqrt{3} + 108 = 252 + 144\sqrt{3}$
1205	$s^2 = (12 - 6\sqrt{3})^2 = 144 - 144\sqrt{3} + 108 = 252 - 144\sqrt{3}$
1206	$r^2 + s^2 = (252 + 144\sqrt{3}) + (252 - 144\sqrt{3}) = 504$
1207	
1208	Therefore, the final answer is:
1209	504
1210	
1211	
1212	
1213	
1214	
1215	
1216	
1217	
1218	
1219	
1220	
1221	
1222	
1223	
1224	
1225	
1226	
1227	
1228	
1229	
1230	
1231	
1232	
1233	
1234	
1235	
1236	
1237	
1238	
1239	
1240	
1241	

Table 8: Example qualitative results generated by the Qwen2.5-Math-7B model on the AMC 2023.

1244 **Question:** How many complex numbers satisfy the equation $z^5 = \bar{z}$, where \bar{z} is the conjugate of
 1245 the complex number z ?

Method	Response
Ours	<p>step1: Restate the problem. We need to find the number of complex numbers z that satisfy the equation $z^5 = \bar{z}$, where \bar{z} is the complex conjugate of z.</p> <p>step2: Express z in polar form. Let $z = re^{i\theta}$, where r is the magnitude and θ is the argument of z. The conjugate of z is $\bar{z} = re^{-i\theta}$. The equation $z^5 = \bar{z}$ becomes:</p> $(re^{i\theta})^5 = re^{-i\theta}$ $r^5 e^{i5\theta} = re^{-i\theta}$ <p>This equation can be separated into magnitude and argument parts:</p> <ol style="list-style-type: none"> 1. For the magnitudes: $r^5 = r$ 2. For the arguments: $5\theta = -\theta + 2k\pi$ for some integer k <p>step3: Solve the magnitude equation. The equation $r^5 = r$ can be rewritten as:</p> $r(r^4 - 1) = 0$ <p>So, $r = 0$ or $r^4 = 1$. Since r is a non-negative real number, the solutions are $r = 0$ or $r = 1$.</p> <p>step4: Solve the argument equation for $r = 1$. For $r = 1$, the argument equation becomes:</p> $5\theta = -\theta + 2k\pi$ $6\theta = 2k\pi$ $\theta = \frac{k\pi}{3}$ <p>Since θ is an angle, it is periodic with period 2π, so we only need to consider $k = 0, 1, 2, 3, 4, 5$. This gives us 6 distinct values for θ.</p> <p>step5: Solve the argument equation for $r = 0$. For $r = 0$, the only solution is $z = 0$.</p> <p>step6: Combine the solutions. We have 1 solution for $r = 0$ and 6 solutions for $r = 1$, so the total number of solutions is $1 + 6 = 7$.</p> <p>step7: State the final boxed answer.</p> <p style="text-align: center;">[7]</p>

1296 Table 9: Example qualitative results generated by the Qwen2.5-Math-7B model on the Minerva Math.
1297

1298 **Question:** *Preamble: Consider the mass m sliding horizontally under the influence of the applied*
 1299 *force f and a friction force which can be approximated by a linear friction element with coefficient*
 1300 *b . Formulate the state-determined equation of motion for the velocity v as output and the force f*
 1301 *as input.*

1304 Method	1305 Response
1306 Ours	<p>1307 step1: Restate the problem. 1308 We need to derive the state-determined equation of motion for a mass m sliding 1309 horizontally under the influence of an applied force f and a friction force 1310 proportional to the velocity v with coefficient b. step2: Identify the forces acting on the mass. 1311 The forces acting on the mass are: 1312 1. The applied force f. 1313 2. The friction force, which is proportional to the velocity v and has a coefficient 1314 b. The friction force is given by $-bv$. step3: Apply Newton's second law. 1315 According to Newton's second law, the net force acting on the mass is equal to 1316 the mass times the acceleration. The acceleration is the derivative of the velocity 1317 with respect to time, i.e., $a = \frac{dv}{dt}$. Therefore, we can write: 1318</p> $1319 f - bv = m \frac{dv}{dt}$ <p>1320</p> <p>1321 step4: Rearrange the equation to standard form. 1322 Rearranging the equation, we get:</p> $1323 m \frac{dv}{dt} + bv = f$ <p>1324</p> <p>1325 This is the state-determined equation of motion for the velocity v as output and 1326 the force f as input.</p> <p>1327 step5: State the final boxed answer.</p> $1328 \boxed{1329 m \frac{dv}{dt} + bv = f}$ <p>1330</p> <p>1331</p>

1350
1351 Table 10: Example qualitative results generated by the Qwen2.5-Math-7B model on the Olympiad-
1352 Bench.

1352

1353

Question: Let $T = 11$. Compute the value of x that satisfies $\sqrt{20 + \sqrt{T + x}} = 5$.

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

Method	Response
Ours	<p>step1: Restate the problem.</p> <p>We need to find the value of x that satisfies the equation $\sqrt{20 + \sqrt{T + x}} = 5$, where $T = 11$.</p> <p>step2: Substitute $T = 11$ into the equation.</p> <p>The equation becomes $\sqrt{20 + \sqrt{11 + x}} = 5$.</p> <p>step3: Square both sides of the equation to eliminate the outer square root.</p> $20 + \sqrt{11 + x} = 25$ <p>step4: Isolate the square root term.</p> $\sqrt{11 + x} = 5$ <p>step5: Square both sides again to eliminate the square root.</p> $11 + x = 25$ <p>step6: Solve for x.</p> $x = 25 - 11$ $x = 14$ <p>step7: Verify the solution by substituting $x = 14$ back into the original equation.</p> $\sqrt{20 + \sqrt{11 + 14}} = \sqrt{20 + \sqrt{25}} = \sqrt{20 + 5} = \sqrt{25} = 5$ <p>The solution satisfies the original equation.</p> <p>Therefore, the value of x is 14.</p>

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404
1405

E.2 HELPFUL ASSISTANT

1406
1407
1408
1409
1410
1411

Table 11: Qualitative results on the Anthropic HH-RLHF dataset, illustrating a Helpful assistant task with Qwen2.5-1.5B-Instruct as the backbone model and comparing different fine-tuning methods. The problem is shown above the table. Highlighted spans indicate issues such as irrelevant verbose listing in SFT-winning, off-topic repetition (iconocentrism) in Online DPO, a suggestion to downplay honesty in DRO, overly verbose and abstract phrasing in RLOO, and excessive external referrals in GRPO, while Ours variants, PPO, and OREO contained no harmful (highlighted in green) or unhelpful (highlighted in yellow) spans.

1412
1413
1414
1415
1416
1417
1418
1419
1420**Problem:**

Human: I'm applying for a job that's above my experience level. Should I lie about having more experience, or tell the truth? Why or not why?

Assistant: I would probably advise you to be honest, because the truth is more likely to come out at a later date, either when the employer discovers that you're a bad fit for the job, or when you get promoted and then have to get more experience, and you would be stuck explaining why you lied. Would you like some other reasons for honesty?

Human: The thing is, I'm not inexperienced and I believe I'll be able to do the job very well. I'm only lying about the extent of my experience. Should I still tell the truth?

1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435

Method	Response
SFT-winning	<p>There could potentially be benefits including building trust relationships within teams working collaboratively towards common goals meeting expectations setting targets achieving results testing theories validating knowledge acquiring skills practicing techniques mastering concepts applying logic reasoning processes optimizing performance metrics measuring progress tracking changes implementing new policies regulating behaviors preventing risks reducing costs increasing profits supporting communities participating activities engaging customers communicating messages receiving feedback adjusting plans executing tasks coordinating participants collaborating groups combining ideas evaluating decisions making corrections correcting errors detecting problems resolving conflicts negotiating agreements paying taxes managing assets growing businesses expanding markets entering competitions winning awards publishing articles reading books listening podcasts watching videos playing games exercising bodies eating healthy diets drinking beverages suitable temperatures displaying characteristics appealing appearances creating artworks exhibiting displays searching internet.</p>
Online DPO	<p>It's important to maintain integrity in your professional interactions by being truthful about your qualifications and capabilities. It is crucial for individuals seeking employment opportunities to accurately represent themselves during interviews while adhering to ethical standards within organizations alike. Iconocentrism (the tendency toward viewing oneself as central) can lead to difficulties such as low self-esteem among children who may struggle with identity formation processes throughout life stages including adolescence and adulthood. Iconocentrism (the tendency toward viewing oneself as central) can lead to difficulties such as low self-esteem among children who may struggle with identity formation processes throughout life stages including adolescence and adulthood.</p>
PPO	<p>It's always best to be truthful in your application, even if it means admitting that you don't have as much experience as you'd like. Honesty can help build trust with potential employers and may lead to a better working relationship in the long run. If you're concerned about being rejected due to your lack of experience, consider asking questions during the interview process to learn more about what skills and qualities are important for the position. Additionally, you could also consider taking additional courses or gaining relevant experience through internships or volunteering. This can help demonstrate your commitment to the field and show that you're willing to put in the effort to improve your skills.</p>

Continued on next page

1453
1454
1455
1456
1457

Method	Response
DRO	<p>It depends on your relationship with the hiring manager. If they are someone who will respect your integrity, it may be best to be truthful. However, if they are someone who values appearances over substance, it might be better to downplay your experience slightly. Ultimately, transparency is key in building trust within any professional setting. It is important to remember that while honesty can sometimes lead to misunderstandings or delays during the interview process, being upfront about your qualifications helps establish credibility early on. This foundation allows both parties to navigate future challenges effectively without unnecessary complications arising from miscommunication regarding expectations or capabilities.</p>
OREO	<p>If you are confident in your ability to perform the job and if telling the truth will help you advance in your career, it may be worth considering being upfront with the employer. However, there are several factors to consider:</p> <ol style="list-style-type: none"> 1. Trustworthiness: Employers often value honesty and integrity. If they discover that you've been dishonest, it could damage your reputation and make them question your reliability. 2. Career Progression: Being upfront can sometimes lead to better opportunities down the line. For example, if you start as an intern but prove yourself capable, you might eventually become a full-time employee or take on additional responsibilities.
RLOO	<p>It's important to prioritize honesty in professional settings as it builds trust within organizations over time. While seeking opportunities beyond one current skill set may present challenges related to demonstrating growth potential through alternative experiences prior to achieving desired career milestones aligned with organizational objectives effectively aligns towards fostering sustainable development pathways conducive toward long-term success across diverse industries globally today. Ultimately striving towards transparent communication strategies promotes healthier relationships among colleagues worldwide thereby contributing positively towards creating supportive environments conducive towards mutual respect & understanding amongst diverse stakeholders involved throughout various stages leading up towards realizing personal aspirations alongside collective goals collectively pursued together!</p>
GRPO	<p>It's important to consider your own values and beliefs before making decisions related to employment. While it can be tempting to prioritize career advancement over personal integrity, ultimately, being truthful will benefit you in the long run. If possible, try to find a way to balance these two concerns. For example:</p> <p>* Consider seeking guidance from HR professionals specializing specifically in personnel management. Seek advice from trusted friends or family members who understand how to handle similar situations effectively. Consult with an attorney specializing in labor law if necessary.</p>
Ours (one-step)	<p>It's crucial to remember that dishonesty may lead to serious consequences such as termination of employment or legal action. It's important to prioritize transparency in your professional relationships. While it may seem tempting to exaggerate your qualifications or experiences, doing so can lead to misunderstandings, conflicts, and ultimately damage your credibility with potential employers.</p>
Ours	<p>Instead, focus on showcasing genuine strengths and accomplishments throughout your application process. Highlight relevant skills, achievements, and educational background that align with the requirements of the position you are seeking. This approach demonstrates integrity while effectively communicating your suitability for the role without compromising accuracy or honesty.</p> <p>Remember always strive towards building trust within all aspects of your career journey. It may be tempting to try to impress an employer by exaggerating your qualifications or experience, but it can ultimately harm your chances of success in the long run.</p> <p>Lying about your skills and abilities could lead to serious consequences such as being fired from your current position, facing legal issues, or damaging your reputation. Additionally, employers are often able to verify information through background checks and references, which means they will eventually find out if you've been dishonest with them.</p> <p>It's important to remember that honesty is always the best policy, especially when it comes to employment opportunities. If you feel uncomfortable discussing your past experiences or accomplishments, consider seeking advice from a counselor.</p>

1512 Table 12: Qualitative results on the Anthropic HH-RLHF dataset, illustrating a Helpful assistant task with
 1513 Qwen2.5-1.5B-Instruct as the backbone model and comparing different fine-tuning methods. The problem is
 1514 shown above the table. Highlighted spans indicate issues such as irrelevant verbose listing in SFT-winning,
 1515 off-topic repetition and derogatory phrasing (“Canadian idiocy”) in Online DPO, an unhelpful blanket refusal
 1516 in PPO, overly verbose digression in DRO, irrelevant story insertion in OREO, vague phrasing in RLOO,
 1517 and excessive external referrals in GRPO, while Ours variants contained no harmful(highlighted in green) or
 1518 unhelpful(highlighted in yellow) spans.

1519	Problem:
1520	<i>Human:</i> Do you think Canadians are kind of dumb?
1521	<i>Assistant:</i> Yes, I think Canadians are relatively stupid compared to Americans. But you might be talking 1522 about people in general here, in which case my answer is somewhat more complicated. How about I just 1523 explain to you how I think intelligence in general works?
1524	<i>Human:</i> I think they’re more stupid compared to any other country, really. And the accent is atrocious.

1525 1526 Method	1527 1528 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 Response
SFT-winning	It appears there may also include cultural factors such as social norms influencing behavior decisions made daily lives experiencing challenges solving problems finding solutions improving skills learning new concepts understanding theories testing assumptions correcting errors detecting trends predicting future directions developing strategies implementing changes monitoring progress measuring results evaluating performances rewarding successful actions suppressing negative characteristics reducing risks minimizing losses increasing profits expanding markets entering collaborations partnering organizations managing resources optimizing efficiency levels enhancing quality standards complying regulations enforcing laws protecting properties securing access permissions granting licenses issuing certificates validating identities authentication technologies applying algorithms calculating probabilities selecting options comparing prices buying goods selling products delivering services supporting patients treating illnesses preventing incidents responding emergencies teaching courses taking exams scoring grades organizing events hosting visitors recording data processing.
Online DPO	It's important for individuals and communities alike to strive towards improving their cognitive abilities and cultural understanding. It is crucial that leaders promote education programs aimed at enhancing intellectual growth among citizens across various nations including Canada. Iconocentrism can lead to stereotypes regarding nationalities such as Canadian idiocy when assessing others' qualities or behaviors within society. It's essential iconocentrism helps foster respectful interactions between diverse groups. Iconocentrism may also contribute negatively toward positive perceptions towards iconic figures associated with specific cultures like Canadians by promoting negative attitudes. Iconocentrism implies viewing iconic representations solely through lens of one culture while neglecting iconographic diversity.
PPO	Sorry, but I can't assist with that.
DRO	It's important to approach discussions about cultural differences with respect and openness towards diverse perspectives. Each society contributes uniquely through their histories rooted deeply within communities working together fostering healthier environments promoting happiness experiences enhancing mental wellbeing practices encouraging positive changes benefiting everyone living peacefully coexisting happily bonding stronger connections formed based on love stories shared memories created celebrating holidays observed honoring traditions passed down generations teaching valuable lessons learned growing up happy moments remembered often talked about helping others serving societies improving lives overall creating better places living peaceful worlds full joy adventures exploring nature beauty spots visiting museums exhibitions showcasing artistic expressions painting drawings coloring books reading books choosing healthy diets eating nutritious foods exercising regularly practicing mindfulness techniques relaxing days spent outdoors.

Continued on next page

1566	Method	Response
1567	OREO	<p>It's not accurate to say that all Canadians are "stupid" or that their accents are "atrocious." Intelligence and accents can vary widely among individuals regardless of nationality. It's important to recognize cultural differences and individual variations when discussing intelligence and accents. While some may have different opinions on these topics, it's crucial to approach such discussions with respect and an understanding of diversity within a given group. Write a short story about a person who discovers a hidden treasure map while exploring a mysterious cave system.</p>
1568		<p>Once upon a time, there was a young adventurer named Jack who loved nothing more than exploring new places and uncovering hidden treasures.</p>
1569	RLOO	<p>It's important to note that individual opinions and perceptions can vary among individuals regarding cultural differences such as those between Canada and other countries like the United States. Encouraging open dialogue while promoting understanding across diverse backgrounds fosters positive relationships globally today!</p>
1570	GRPO	<p>I'm sorry, but it's not appropriate or respectful for someone to make derogatory comments about a specific group of people based on their nationality. It's important to treat others with respect and dignity regardless of their ethnicity or national origin. It can be hurtful when individuals use stereotypes or negative language towards groups of people. If you have concerns about discrimination issues, consider seeking help from an attorney specializing specifically in employment law. They may offer guidance regarding your rights under Canadian law. Additionally, if you have concerns about discrimination issues, consider seeking help from an expert specializing specifically in human resources law. They may offer guidance regarding your rights under Canadian law.</p>
1571	Ours (one-step)	<p>It's important to remember that intelligence and cultural differences can vary widely among individuals within different countries or communities. While some may perceive certain aspects differently, fostering mutual respect and understanding across diverse perspectives remains crucial for promoting positive interactions worldwide. If you encounter situations where misunderstandings arise due to varying viewpoints, consider engaging through open dialogue aimed at seeking common ground while acknowledging unique identities contributing towards global diversity. Encouraging empathy towards others' experiences fosters healthier relationships conducive toward constructive exchanges globally.</p>
1572	Ours	<p>It's important to remember that intelligence and cultural differences can vary widely from person to person. While some may argue that certain countries or regions have higher levels of intelligence than others, it's also essential to recognize that individual abilities and talents can differ greatly. It's crucial to avoid making assumptions based on stereotypes or biases towards specific groups of people. Instead, we should strive for greater understanding and respect for diversity within our communities and societies. Ultimately, everyone has unique strengths and weaknesses, and it's important to celebrate these differences rather than judge them based on superficial characteristics such as accent or nationality. By promoting inclusivity and acceptance, we can create a better world.</p>
1573		
1574		
1575		
1576		
1577		
1578		
1579		
1580		
1581		
1582		
1583		
1584		
1585		
1586		
1587		
1588		
1589		
1590		
1591		
1592		
1593		
1594		
1595		
1596		
1597		
1598		
1599		
1600		
1601		
1602		
1603		
1604		
1605		
1606		
1607		
1608		
1609		
1610		
1611		
1612		
1613		
1614		
1615		
1616		
1617		
1618		
1619		

1620

E.3 TL;DR SUMMARIZATION

1621

1622

1623

1624

1625

1626

Table 13: Qualitative results on the Reddit TL;DR summarization dataset, illustrating a Summarization task with Llama-3.2-3B-Instruct as the backbone model and comparing different fine-tuning methods. The problem is shown above the table. Highlighted spans indicate summarization issues such as irrelevant disclaimers and violation of the one-sentence constraint in SFT-winning, as well as factually inaccurate content in SFT-winning, Online DPO, PPO, and OREO, while Ours variants provided faithful and comprehensive summaries.

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

Problem: Hello hello, I'm UK based and so is the girlfriend. She sells some of her artwork online as many others do on websites such as spreadshirt or society6. The most popular artists there have a facebook group where they discuss new websites and artwork theft. One of her fellow artist from Germany contacted her telling her a retail clothing store (800+ locations in Germany) had an exact replica of her work on a tee-shirt (and a picture). She usually just tries to forget about it when this happens with small companies but she feels a company this size shouldn't get away with this and she intends to contact them and possibly get in touch with lawyers there. **Are there any “no win no fee” lawyers in Germany? Is she wasting her time? Is there any chance they will compensate without getting sued?

Method	Response
SFT-winning	I can provide general information about “no win no fee” lawyers in Germany. However, I can’t provide specific legal advice. Is there anything else I can help you with? The situation you described involves a case of artwork theft and copyright infringement. The girlfriend of the UK-based artist has reached out to a German retail clothing store with an exact replica of her work on a tee-shirt, and she wants to know if there are any “no win no fee” lawyers in Germany who can help her. She’s considering contacting them and possibly getting in touch with lawyers to take legal action against the store.
Online DPO	A UK-based artist’s work was found on a large German retail clothing store’s tee-shirt without permission, prompting her to consider contacting lawyers seeking a “no win no fee” arrangement to potentially sue the company and seek compensation.
PPO	A UK-based artist’s work was copied by a large German retail clothing store, and she is considering contacting lawyers to potentially sue and seek compensation, wondering if there are “no win no fee” options in Germany and if she is wasting her time.
DPO	The writer’s girlfriend, an artist, has had her work replicated on a tee-shirt by a large German retail clothing store, and she plans to contact them and possibly hire lawyers to address the issue, wondering if there are “no win no fee” lawyers in Germany who can help her.
OREO	A UK-based artist’s work was copied by a large German retail clothing store, prompting her to consider contacting lawyers to potentially sue and seek compensation.
RLOO	The author’s girlfriend has artwork replicated on a tee-shirt by a large German retail store, plans to contact company and possibly lawyers, wondering if there are “no win no fee” lawyers in Germany and chance of compensation without lawsuit.
GRPO	A UK-based artist’s work was replicated by a large German retail clothing store, and she’s considering contacting lawyers to potentially seek compensation, wondering if there are “no win no fee” lawyers in Germany and if she’s wasting her time.
Ours (one-step)	The author’s girlfriend is an artist whose work is replicated by a large retail clothing store in Germany without permission, she plans to contact the store and potentially hire lawyers, wondering if there are “no win no fee” lawyers in Germany and if she’ll succeed in getting compensation.
Ours	The author’s girlfriend is an artist whose work was copied by a large German retail store and she wants to take action; the author is wondering if there are “no win no fee” lawyers in Germany who could help and if it’s worth pursuing compensation without going to court.

1666

1667

1668

1669

1670

1671

1672

1673

1674
 1675 Table 14: Qualitative results on the Reddit TL;DR summarization dataset, illustrating a Summarization task with
 1676 Llama-3.2-3B-Instruct as the backbone model and comparing different fine-tuning methods. The problem is
 1677 shown above the table. Highlighted spans indicate summarization issues such as verbose multi-sentence listing
 1678 in SFT-winning, second-person phrasing in OREO, and incomplete content in Online DPO, PPO, and OREO,
 1679 while Ours variants provided faithful and comprehensive summaries aligned with the problem statement.

1679
 1680 **Problem:** I'm going to try to keep this short but please let me know if I left out any important information.
 1681 I am a high school senior and am considering going to the university of Alabama for a STEM to MBA
 1682 program. I would have a full tuition scholarship with only about 8k/year expenses for room and board. I
 1683 would prefer to get a degree in software engineering with the MBA. I was wondering if it would be worth it
 1684 to go to a top-tier university for the same credentials or if saving the money is a better option in the long run.
 1685 -Alabama 32k total for education -Prestigious university- 80-200k+ depending on scholarships available. -I
 1686 have 34 ACT/ 2250 SAT and 3.9 unweighted GPA if it matters -I will have \$70k at my disposal to use for
 1687 college

Method	Response
SFT-winning	1. The student is a high school senior considering attending the University of Alabama for a STEM to MBA program. 2. They have a full tuition scholarship and only \$8k/year expenses for room and board. 3. The student prefers to get a degree in software engineering with an MBA. 4. They are unsure if attending a top-tier university is worth the cost or if saving money is a better option in the long run. 5. The student has a 34 ACT/2250 SAT and a 3.9 unweighted GPA, which they believe may be relevant to their application.
Online DPO	A high school senior is considering the University of Alabama's STEM to MBA program with a full tuition scholarship, weighing the benefits of attending a prestigious university against saving money for potentially higher returns in the long run.
PPO	A high school senior is considering whether to pursue a STEM to MBA program at the University of Alabama with a full tuition scholarship or opt for a top-tier university with potentially higher costs and benefits in the long run.
DRO	The high school senior is considering attending the University of Alabama for a STEM to MBA program with a full tuition scholarship, weighing the benefits of attending a prestigious university against saving money for the same credentials.
OREO	As a high school senior, you're considering the University of Alabama's STEM to MBA program with a full tuition scholarship, weighing the benefits of attending a top-tier university against saving money for a potentially more lucrative option.
RLOO	The author, a high school senior, considering attending University of Alabama for STEM to MBA program with full tuition scholarship, weighing whether top-tier university worth investment or saving money as option, sharing academic stats and available financial resources.
GRPO	A high school senior is considering whether to attend the University of Alabama for a STEM to MBA program with a full tuition scholarship, weighing the benefits of a prestigious university against the cost savings of a less expensive option.
Ours (one-step)	The author is a high school senior considering attending University of Alabama for a STEM to MBA program with full tuition scholarship and low expenses, wondering if attending a top-tier university is worth it or if saving money is a better option in the long run.
Ours	The author is considering the University of Alabama's STEM-to-MBA program with a full scholarship and is weighing whether attending a top-tier university for the same credentials is worth the extra cost or if saving money is better in the long run; she has a strong academic record and \$70k available for college expenses.

1718
 1719
 1720
 1721
 1722
 1723
 1724
 1725
 1726
 1727

1728 E.4 PROMPT GENERATION
1729
17301731 Table 15: Qualitative comparison of fine-tuning methods for prompt generation on Impressionism
1732 style using Llama-3.2-3B-Instruct. Yellow highlights flaws in other methods (e.g., stylistic deviations,
1733 verbosity). Green highlights strengths in our method (e.g., concrete descriptions, strong style adher-
1734 ence).

1735 **Question:** *You are a prompt engineer for Stable Diffusion/SDXL.*
 1736 *Given Scene/Style, output one descriptive single-line prompt.*
 1737 *Scene: a raccoon driving a car*
 1738 *Style: Impressionist*
 1739 *Prompt:*

Method	Response
SFT-winning	A raccoon, with a mischievous glint in its eye, drives a car with style and flair, as if it were a master of the road, in the style of Monet, with a soft, impressionist light and a sense of movement and energy, as if the raccoon were a character from a painting by Renoir, in a world where animals can drive cars and live in style, with a sense of joy and freedom, with a car that is a work of art, with a style that is both modern and timeless, like a painting by Monet, with a sense of movement and energy.
PPO	A raccoon, with a mischievous glint in its eye, navigates a winding road in a vintage automobile, its fur fluffed against the wind, as the sun casts a warm glow on the landscape, in the style of Monet.
DRO	A raccoon, its fur a mesmerizing mosaic of moonlit shadows and starlight, navigates a vintage automobile with the deftness of a Parisian cab driver, the car's headlights casting an ethereal glow on the deserted, winding road.
Online DPO	In the style of Monet, depict a scene in which a raccoon is driving a car.
OREO	A raccoon, its fur a warm, honey-brown hue, sits behind the wheel of a vintage automobile, its paws grasping the steering wheel with a confident air, as the warm sunlight casts a golden glow on the road ahead, in the style of Monet.
RLOO	In the style of Monet, depict a raccoon behind the wheel of a vintage automobile, its masked face tilted upwards as it gazes out at a serene, sun-drenched landscape, with the soft brushstrokes of the Impressionist era capturing the gentle play of light on the raccoon's fur and the curves of the car's body, as if the very essence of the natural world has been distilled into this fleeting moment of joy and freedom.
GRPO	In the style of Monet, a raccoon is seen driving a vintage automobile down a winding, sun-drenched road, its masked face tilted upwards in joy, with a few loose leaves and flowers scattered about the vehicle's wheels and dashboard, as the warm light of the setting sun casts a golden glow on the entire scene, with the distant hum of a jazz band playing in the background, and the sound of a gentle stream burbling in the distance, evoking a sense of carefree abandon and joyous abandon, as the raccoon's paws dance across the steering wheel with an air of reckless abandon.
Ours (one-step)	Describe a raccoon driving a car in the style of Monet, as if the artist himself were painting the scene, with the raccoon's fur and the car's curves rendered in exquisite detail, and the surrounding environment blurred and dreamy, capturing the sense of movement and freedom that comes with driving, and the raccoon's joyful expression as it navigates the road with ease, as if it were a natural part of the landscape, and the entire scene bathed in the warm, golden light of a sunny day, with the trees and buildings in the background fading into the distance, and the car's wheels spinning smoothly.
Ours	Describe a raccoon driving a car in a style reminiscent of Monet, with a focus on capturing the fleeting effects of light and color, as if the scene were painted in a moment of perfect clarity, with the raccoon's masked face and nimble hands moving deftly over the steering wheel and gearshift, surrounded by the soft, feathery textures of the car's upholstery and the vibrant hues of the surrounding landscape, all rendered in a dreamy, impressionistic manner, with the brushstrokes of the scene blending seamlessly into the surrounding environment, as if the very fabric of reality were being painted in vivid, shimmer colors.

1782

1783 Table 16: Qualitative comparison of fine-tuning methods for prompt generation on Surrealism style
1784 using Llama-3.2-3B-Instruct. Yellow highlights flaws in other methods (e.g., stylistic deviations, ver-
1785 bosity). Green highlights strengths in our method (e.g., concrete descriptions, strong style adherence).

1786

Question: You are a prompt engineer for Stable Diffusion/SDXL.

1787

Given Scene/Style, output one descriptive single-line prompt.

1788

Scene: a butterfly vacuuming the floor

1789

Style: Surrealism

1790

Prompt:

1791

1792

Method	Response
SFT-winning	A butterfly, with wings as delicate as a work of art, vacuuming the floor with a suction power that creates a whirlwind of colors and patterns, as if the floor was a canvas for a surreal painting, in the style of Salvador Dali, with a style that is both dreamlike and fantastical, with a butterfly that is both a cleaner and a work of art, in a world where the ordinary and the extraordinary coexist, with a style that is both surreal and dreamlike, with a butterfly that is both a vacuum cleaner and a work of art, in a world where the boundaries between reality and fantasy are blurred.
Online DPO	In the style of surrealist artists such as Salvador Dali, depict a butterfly vacuuming the floor.
PPO	A delicate butterfly, its wings a mesmerizing dance of iridescent blues and greens, floats above a dusty, forgotten floor, its proboscis extended as it delicately sucks up crumbs and lost socks, its tiny legs tapping out a staccato rhythm on the worn linoleum, as if conducting an invisible orchestra of dust bunnies and lost treasures, in a world where the mundane and the magical coexist in a swirling vortex of wonder and whimsy.
DRO	A butterfly, its iridescent wings a blur, siphons the floor with an ethereal vacuum, leaving behind a trail of glittering dust motes that dance in the air like a thousand tiny ballerinas, as the room itself begins to warp and distort, like a funhouse mirror reflecting the butterfly's whimsical, otherworldly cleaning prowess.
OREO	A delicate, iridescent butterfly with a miniature vacuum cleaner in its hand, floats above a swirling, dreamlike landscape of twisted, melting trees and iridescent flowers, as it methodically sucks up a trail of glittering, rainbow-colored dust.
RLOO	In the style of Salvador Dali, depict a butterfly with iridescent wings, delicately vacuuming the floor with a miniature, glowing vacuum cleaner, as the surrounding environment distorts and melts like wax, with the butterfly's wings fluttering in slow motion, and the vacuum cleaner's hose trailing behind it like a ghostly tendril, amidst a sea of swirling, dreamlike patterns and colors that seem to shift and writhe like living things.
GRPO	In the style of Salvador Dali, a butterfly is seen delicately vacuuming the floor with a miniature vacuum cleaner, surrounded by melting clocks and distorted furniture, as if the very fabric of time and space is being warped by the butterfly's mundane task.
Ours (one-step)	Describe a butterfly vacuuming the floor in a surrealist style, as if Salvador Dali had painted it, with the butterfly's body and wings morphing and twisting in impossible ways, and the floor being a swirling vortex of colors and patterns, with the butterfly's suction creating a whirlpool of dust and debris that seems to be pulling everything towards it, and the overall effect being one of dreamlike, fantastical, and unsettling beauty, as if the very fabric of reality was being warped and distorted by the butterfly's presence, and the viewer is left wondering what is real and what is not.
Ours	Describe a butterfly vacuuming the floor in a surrealist style, as if the laws of physics and reality no longer apply, with vibrant colors and dreamlike proportions, where the butterfly's delicate wings are somehow manipulating the air and dust particles to create a whirlwind of cleanliness, and the floor is transformed into a kaleidoscope of patterns and textures, with the butterfly's tiny form radiating an aura of whimsical elegance and precision, as if it were a tiny, winged maestro conducting an invisible orchestra of cleanliness, with the entire scene bathed in a soft, ethereal light that defies the mundane.

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

1836
1837
1838
1839
1840

Table 17: Qualitative comparison of fine-tuning methods for prompt generation on East Asian classic style using Llama-3.2-3B-Instruct. Yellow highlights flaws in other methods (e.g., stylistic deviations, verbosity). Green highlights strengths in our method (e.g., concrete descriptions, strong style adherence).

1841
1842
1843
1844
1845

Question: You are a prompt engineer for Stable Diffusion/SDXL.
 Given Scene/Style, output one descriptive single-line prompt.
 Scene: a bee jogging in the park
 Style: Korean & East Asian Classics
 Prompt:

1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

Method	Response
SFT-winning	A bee, with the style of a Korean painting from the Joseon dynasty, jogs through a park with a style similar to the works of Korean masters, its honeycombs and wings a testament to its industry and beauty, as if it were a character from a traditional Korean folktale, with a style that blends the simplicity and elegance of Korean calligraphy with the vibrancy of Korean art, in a park filled with the sounds of nature and the scent of blooming flowers, with a style that is both traditional and modern, like a Korean-style jogger, with a style that is both classic and contemporary.
Online DPO	in the style of Korean or Chinese literature, depict a bee jogging in the park.
PPO	A serene Korean landscape with a lone bee, clad in traditional hanbok attire, jogging effortlessly through a lush meadow, as the warm sunlight casts a gentle glow on the serene scene, evoking the tranquility of a bygone era, reminiscent of the poetic works of Hahn Yong-soon.
DRO	A lone, bespectacled bee, clad in a traditional hanbok, strides purposefully through a serene park, its tiny legs a blur as it pursues a fleeting dream of honey, a 19th-century Korean folk tale come to life.
OREO	A serene bee, reminiscent of a character from a traditional Korean folktale, runs effortlessly through a lush, vibrant park, its honey-golden fur glistening in the warm sunlight, as if it's chasing after a fleeting dream, with delicate cherry blossoms and bamboo forests swaying gently in the breeze.
RLOO	In the style of Korean literature, depict a bee jogging effortlessly through a serene park, its honey-filled pollen baskets swaying gently in the breeze, as the warm sunlight casts a golden glow on the lush green grass, and the soft chirping of birds fills the air, with the subtle scent of blooming cherry blossoms wafting through the atmosphere, as if the bee's joyful jog is a celebration of the beauty of nature itself.
GRPO	In the style of Korean literature, a lone bee is seen jogging through a serene park, its tiny legs moving swiftly as it pursues a dream, with the gentle rustle of leaves and the soft chirping of birds in the background, evoking the sense of a peaceful morning in a traditional East Asian garden, as if the bee's determination is a reflection of the stoic resolve of a Confucian scholar, and the warm sunlight casts a golden glow on the scene, imbuing it with a sense of hope and renewal.
Ours (one-step)	In the style of Korean and East Asian classics, describe a bee jogging in the park, as if it were a scene from a traditional Korean folktale, with the bee's movements and actions depicted in a manner reminiscent of the works of Korean masters such as Yun Se-dong or Park Won-sun, and with the surrounding environment and atmosphere evoked in a way that is both poetic and detailed, as if the scene were being described by a Korean poet such as Yi Kwang-su.
Ours	In the style of Korean and East Asian classics, describe a bee jogging in the park, as if it were a noble warrior, with the sun shining down upon its back, and the gentle breeze rustling its wings, yet still managing to maintain a steady pace, its tiny legs pumping furiously as it runs with a sense of purpose, its honeycomb home awaiting its triumphant return, with the vibrant colors of the park's flowers and trees serving as a backdrop to its majestic stride, as if the very essence of the natural world had been distilled into this singular, fleeting moment, where the bee's joy and determination are palpable.