# Enhancing Diversity in Large Language Models via Determinantal Point Processes

**Yilei Chen**[*] **Souradip Chakraborty**[†]

**Lorenz Wolf**[‡] **Ioannis Ch. Paschalidis**[*] **Aldo Pacchiano**[*§]

## Abstract

Reinforcement learning (RL) has emerged as a popular method for post-training large language models (LLMs). While improving the model's performance on downstream tasks, it often reduces the model's output diversity, leading to narrow, canonical responses. Existing methods to enhance diversity are limited, either by operating at inference time or by focusing on surface-level differences. We propose a novel training method named DQO (**D**iversity **Q**uality **O**ptimization) based on determinantal point processes (DPPs) to jointly optimize LLMs for quality and semantic diversity. Our approach samples and embeds a group of responses for each prompt, then uses the determinant of a kernel-based similarity matrix to measure diversity as the volume spanned by the embeddings of these responses. DQO is flexible and can be applied on top of existing RL algorithms. Experiments across instruction-following, summarization, story generation, and reasoning tasks demonstrate that our method substantially improves semantic diversity without sacrificing model quality.

## 1 Introduction

Large language models (LLMs) are typically post-trained to better align with human intentions and to perform effectively on downstream tasks [Ouyang et al., 2022, Bai et al., 2022]. Reinforcement learning (RL) is commonly used to either maximize an existing reward function, or a reward model trained from human preference data [Ziegler et al., 2020, Stiennon et al., 2020, Ouyang et al., 2022, Bai et al., 2022, DeepSeek-AI et al., 2025]. These methods substantially improve the output quality for targeted tasks. However, a widely observed drawback is that post-training often leads to a sharp reduction in output diversity, with models converging on a narrow set of canonical responses [Kirk et al., 2023, Murthy et al., 2024, Anderson et al., 2024, Xu et al., 2025, Casper et al., 2023]. This loss of diversity is problematic across multiple dimensions: it limits reasoning and personalization by restricting alternative solution paths or user-preferred styles; it undermines test-time performance by reducing test-time search capabilities, robustness to distribution shift, and coverage of reward modes; and it weakens training dynamics by limiting exploration and the discovery of novel strategies.

Current efforts to promote diversity in LLM outputs are mostly limited to inference-time interventions such as temperature scaling [Ackley et al., 1985], top-k sampling [Holtzman et al., 2020], and related strategies [Nguyen et al., 2024, Franceschelli and Musolesi, 2025]. While these approaches help in improving the spread across the response distribution, they often only introduce shallow, token-level variation and fail to produce truly diverse or meaningful responses. In many cases, they also hurt

[*]Boston University
[†]University of Maryland
[‡]University College London
[§]Broad Institute of MIT and Harvard

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

output quality, making the model less reliable or coherent. More critically, they cannot recover modes absent from the base model's learned distribution. A more principled approach is to optimize for diversity during training, but this poses two central challenges: defining diversity in a computationally efficient and theoretically sound way, and balancing diversity with response quality.

Recent attempts to improve diversity during training largely remain at the lexical level. Yao et al. [2025] encourage variation with a token-level entropy regularizer, but such measures fail to capture semantic diversity, which is often more meaningful to humans. Lanchantin et al. [2025] extend direct preference optimization (DPO) [Rafailov et al., 2023] by selecting the most diverse candidate among high-reward responses, yet their notion of diversity is still based on surface features such as generation probability or word counts. Likewise, Li et al. [2025] aim to preserve diversity in supervised fine-tuning by carefully constraining probability transfer between tokens during updates, again focusing on token-level variation. More broadly, these approaches prioritize local lexical differences rather than encouraging models to generate responses that span distinct semantic modes. Most related to our work, Chung et al. [2025] introduce a DPO variant that weights loss by average embedding distance, but it remains DPO-specific and considers only pairwise distances, which can yield degenerate solutions.

In this work, we propose a principled training method based on determinantal point processes (DPPs) [Kulesza et al., 2012] to directly optimize LLMs for both quality and diversity in generated responses. Unlike token-level entropy or lexical perturbations, our approach operates at the semantic level. Specifically, for each prompt we sample a set of responses, map them into an embedding space using a pretrained encoder, and compute a similarity matrix via a kernel function. The diversity score is then defined as the determinant of this matrix, which corresponds to the volume spanned by the response embeddings. Optimizing this objective encourages the model to generate responses that span a subspace in the answers' embedding space with the largest volume. The reward of each response can be regarded as a scaling factor of the corresponding embedding vector, providing an interpretable mechanism to balance quality against diversity. We refer to our algorithm as DQO (**D**iversity **Q**uality **O**ptimization). DQO is highly flexible and can be layered on top of existing state-of-the-art methods such as GRPO, making it broadly applicable in practice. We evaluate DQO across instruction-following, summarization, story generation, and reasoning tasks, and demonstrate that it significantly enhances semantic diversity while maintaining high response quality.

We summarize our contributions as the following,

- **Principled framework for Diversity Quality Optimization** : We propose a principled method, DQO, for post-training LLMs to generate diverse, high-quality responses. DQO is a flexible approach that can be applied on top of existing reinforcement learning algorithms, such as PPO and GRPO.
- **Semantic diversity beyond lexical variation** : We demonstrate that the DPP-based formulation provides a theoretically grounded framework for defining diversity, ensuring that responses span the semantic space both broadly and meaningfully.
- **Quality–diversity trade-off** : We experimentally show that DQO improves semantic diversity while preserving response utility, coherence, and task accuracy across a wide range of tasks. In addition, we conduct extensive ablations to illustrate the trade-off between quality and diversity.

## 2 Related works

**Evaluating Diversity of LLMs.** Several works have focused on evaluating the diversity of LLM generated content [Guo et al., 2024, Shaib et al., 2024], also on investigating the impact of post-training on diversity metrics [Kirk et al., 2023, Shypula et al., 2025]. The lack of diversity in LLM generated content also affects text written by humans using LLMs [Padmakumar and He, 2024].

**Improving Diversity of LLMs.** There are mainly two lines of works on promoting diversity in LLMs. One focuses on inference strategies. Nguyen et al. [2024] proposed a decoding method to reallocate the next-token probabilities which they show can increase the entropy of the correct solutions. The DiffSampling strategy, proposed by Franceschelli and Musolesi [2025], considers the largest difference between consecutive probabilities of tokens in a sorted distribution to promote diversity while maintaining correctness. Ahmed et al. [2025] proposed a two-stage inference strategy which consists of a high-temperature key words sampling process and a low-temperature expansion procedure.

Another line of works focus on the training strategy to best elicit diversity from LLMs. Lanchantin et al. [2025] proposed diverse preference optimization. They selected the most diverse response from the high-reward group and the least diverse response from the low-reward group to form the preference pair. The selection is based on some diversity criteria. Yao et al. [2025] shows that by adding an entropy term of correct answers to the reward-based objective, LLMs can improve the diversity while maintaining the quality. Different from those using reinforcement learning algorithms, Li et al. [2025] instead study the supervised finetuning approach. They proposed carefully-designed update strategy to mitigate the distribution collapse in SFT, thus encourages diversity. Most related to our work, Chung et al. [2025] propose a variant of DPO that weights the loss by the average pairwise distance in cosine similarity after embedding responses, this however, is limited to DPO, considers only pairwise distances, and requires sampling $k \geq 3$ responses per prompt in the training dataset.

**Determinantal Point Processes.** Determinantal point processes (DPPs) [Kulesza et al., 2012], are a class of probabilistic models that arise in quantum physics and random matrix theory for modeling repulsion. DPPs are well-suited for modeling diversity. Parker-Holder et al. [2020] proposed a DPPs-based algorithm to train a population of diverse polices in reinforcement learning for better exploration.

# 3 Preliminaries

## 3.1 Notations

For ease of readability, we summarize some frequently used notations here. We use $x$ and $y$ to represent a prompt and a response, respectively. We represent a group of $k$ responses $\{y_1, \ldots, y_k\}$ by $y_{1:k}$ and we denote $\{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_k\}$ by $y_{-i}$. We use $I_k \in \mathbb{R}^{k \times k}$ to represent the identity matrix with size $k$. And $\det(\cdot)$ represents the determinant of a matrix.

## 3.2 Reinforcement learning

Reinforcement learning has become a widely adopted approach for post-training LLMs with either an existing reward function or the one inferred from a preference dataset (e.g., RLHF). With the reward function, the model is typically optimized by maximizing the following KL-regularized objective,

$$\pi^* = \arg\max_{\pi_\theta} \{J(\pi_\theta) - \beta KL(\pi_\theta || \pi_{ref})\} \tag{1}$$

where $J(\pi_\theta) = \mathbb{E}_{x, y \sim \pi(\cdot|x)}[r(x, y)]$ is the expected return and $\beta$ is a hyperparameter that balances the KL divergence penalty and the rewards. Among existing algorithms, PPO [Schulman et al., 2017] and GRPO [Shao et al., 2024] have demonstrated strong empirical performance, by introducing some practical techniques including the clipping mechanism and group-based advantage estimation, respectively.

## 3.3 Determinantal point processes (DPPs)

In this work, we quantify the diversity of LLM generated outputs based on ideas derived from the Determinantal Point Process (DPP) literature (for a comprehensive introduction to DPPs, please refer to Kulesza et al. [2012]). Below we introduce the definition of an L-ensemble, which is a subclass of DPPs.

**Definition 1** (**$L$-ensemble**) *Let $\mathcal{Y} = \{1, 2, \ldots, N\}$ be a ground set, and $\mathbf{Y} \subseteq \mathcal{Y}$ be a random subset. Suppose $L \in \mathbb{R}^{N \times N}$ is a real symmetric positive semi-definite matrix. We say $L$ defines an L-ensemble, if for every $A \subseteq \mathcal{Y}$,*

$$\Pr(\mathbf{Y} = A) \propto \det(L_A),$$

*where $L_A$ is the submatrix of $L$ indexed by $A$.*

If we think of the entries of $L$ as measurements of similarity between pairs of elements, such as the dot product of the feature vectors of items, the determinant $\det(L_A)$ corresponds to the squared

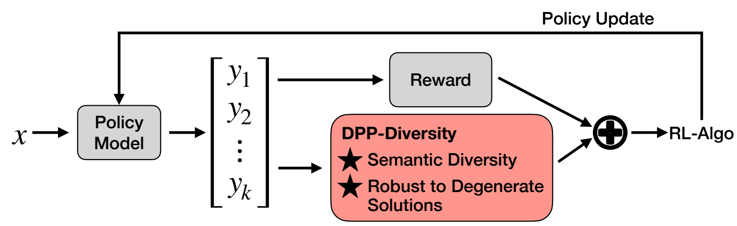



Figure 1: The DPP-based diversity metric of DQO promotes meaningful semantic diversity while ensuring robustness agains degenerate solutions.

volume spanned by the feature vectors of items in $A$, which increases when the vectors are diverse and decreases when they are redundant or highly correlated. Thus, DPPs are well-suited to represent distributions over subsets of items where sets consisting of diverse items are more likely.

For instance, consider the two-element set $A = \{i, j\}$. By definition,

$$\Pr(A) \propto \begin{vmatrix} L_{ii} & L_{ij} \\ L_{ji} & L_{jj} \end{vmatrix} = L_{ii}L_{jj} - L_{ij}L_{ji}$$

If items $i$ and $j$ are highly similar, then $L_{ij}$ is large, resulting in a low probability. Conversely, if $L$ is diagonal (i.e., $L_{ij} = 0$), there are no correlations and the elements occur independently.

# 4 Diversity Quality Optimization

We now formally define our notion of semantic diversity inspired by DPPs and how to incorporate it in reinforcement learning algorithms to jointly optimize quality and diversity.

## 4.1 Diversity via DPPs

Based on the above definition of DPPs, given a group of responses $y_{1:k}$, we can formulate their diversity score as,

$$\mathrm{Div}(y_{1:k}) = \det(L_\phi(y_{1:k})) \tag{2}$$

where $L_\phi(y_{1:k})[i,j] = f(\phi(y_i), \phi(y_j))$, $f$ is a kernel function and $\phi(\cdot)$ is a selected embedding model which can map a response into a high-dimensional semantic space. Although in most of this work we set the kernel function as the dot product, $f(\phi(y_i), \phi(y_j)) = \langle \phi(y_i), \phi(y_j) \rangle$, in Appendix C we also explore Gaussian kernel function and provide additional results. For simplicity, when it is clear from the context, we will omit the subscript in $L_\phi$.

Our definition of diversity offers two main advantages. First, it operates in the embedding space of the responses, allowing it to capture semantic diversity which is typically what humans intend. Second, its determinant-based definition induces a notion of group diversity that overcomes the limitations of simple pairwise distance metrics. A key limitation of pairwise distance measures, such as the average distance across responses, is their well-known vulnerability to a degenerate "clustering" effect, as noted by Parker-Holder et al. [2020]. An algorithm optimizing for this metric might produce responses that form a few distinct, widely separated clusters, creating a misleading sense of diversity. In contrast, our determinant-based metric, which encourages the formation of a parallelepiped with a large volume in the embedding space, directly addresses this issue. The determinant is highly sensitive to the linear independence of the response vectors. If responses form tight clusters, the vectors within a cluster become nearly linearly dependent, causing the determinant of the similarity matrix to approach zero, regardless of the large distances between clusters. This correctly identifies a lack of true diversity. Furthermore, the determinant is a more robust measure because it recognizes when responses, despite having large pairwise distances, are confined to a lower-dimensional subspace. This forces the system to explore the full high-dimensional embedding space, ensuring genuine diversity that simple pairwise distances fail to capture.

### 4.2 Quality-Diversity objective

Using the DPP based diversity metric (2) we now present the objective optimized by our DQO algorithm. For each prompt $x$, we sample $k$ responses $y_{1:k} \sim \pi_\theta(\cdot|x)$ from the policy, similarly to the sampling performed as part of GRPO. Instead of optimizing only the reward, we incorporate a diversity term based on the logarithm of our diversity metric into the objective. This yields the objective given by

$$J_{Div}(\pi_\theta) = \mathbb{E}_{x, y_{1:k} \sim \pi_\theta(\cdot|x)} \left[ \sum_{i=1}^{k} r(x, y_i) + \alpha \log \det(L_\phi(y_{1:k})) - \beta KL(\pi_\theta || \pi_{ref}) \right], \quad (3)$$

where $L_\phi(y_{1:k})$ is defined in (2). The hyperparameter $\alpha$ controls the trade-off between quality and diversity. Maximizing $J_{Div}(\pi_\theta)$ directly optimizes the policy for both quality and semantic diversity in generated responses.

In fact, it can be shown that by optimizing (3), the optimal policy satisfies,

$$\pi_{div}(y_{1:k}|x) \propto \pi_{ref}(y_{1:k}|x) \exp\left(\frac{1}{\beta}\left(\sum_{i=1}^{k} r(x, y_i) + \alpha \log \det(L_\phi(y_{1:k}))\right)\right) \quad (4)$$

For simplicity of exposition, suppose $\beta = \alpha$. We can define a reward-augmented embedding vector for the prompt-response pair $(x, y)$ as $\psi(x, y) = \sqrt{exp\left(\frac{r(x,y)}{\alpha}\right) \pi_{ref}(y|x)} \cdot \phi(y)$. Here, the reward acts as a scaling factor of the original semantic embedding. With the formulation of the reward-augmented embeddings, we can show that our optimal policy satisfies,

$$\pi_{div}(y_{1:k}|x) \propto \det(L_\psi(x, y_{1:k})) \quad (5)$$

For the complete derivation, we refer the reader to Appendix A. The expression above shows that our optimal policy (5) assigns probabilities to groups of responses in proportion to the determinant of the Gram matrix constructed from their embedding vectors. Geometrically, this means the policy selects groups of vectors in the response embedding space according to the squared volume of the parallelepiped spanned by those vectors.

The balance between quality and diversity also admits a clear geometric interpretation. The embedding vector $\psi$ consists of two components: a semantic embedding vector, which determines its direction, and a reward, which determines its norm. To maximize the volume of the spanned space, one should select vectors that are well separated from each other (i.e., diverse responses) while also having large norms (i.e., high-quality responses). The overall trade-off is governed by the hyperparameter $\alpha$.

**Connections with D-Optimal Design.** In the classical theory of experimental design, the goal is to select design points that are most informative for estimating an unknown parameter vector. Consider the linear model $y = z^\top \theta + \xi$, where one collects design points $z_1, \ldots, z_n \in \mathbb{R}^d$. These design points represent the experimental conditions under which data is observed, and their selection determines how precisely the parameter $\theta$ can be estimated. The associated information matrix is defined as

$$M = \sum_{i=1}^{n} z_i z_i^\top = ZZ^\top,$$

where $Z = [z_1, \ldots, z_n]$ is the $d \times n$ design matrix. The criterion of *D-optimal design* selects these design points to maximize $\det(M)$, since this minimizes the volume of the confidence ellipsoid for the unknown parameter $\theta$ [Kiefer, 1959, Pukelsheim, 2006]. By Sylvester's determinant identity,

$$\det(M) = \det(ZZ^\top) = \det(Z^\top Z),$$

(see, e.g., Horn and Johnson [2012]), so D-optimality is equivalently expressed as maximizing the determinant of the *Gram matrix* (the kernel matrix) of the design vectors under the dot-product kernel. Our DQO objective can be seen as a direct analogue of this construction. The role of the design vectors $z_i$ is instead played by the reward-augmented embeddings $\psi(x, y)$, which incorporate both the semantic content of a response and its quality signal. Maximizing the determinant of the Gram matrix built from these embeddings is therefore analogous to maximizing information gain in D-optimal design: it encourages the selected responses to be as linearly independent as possible in the $\psi$-space, ensuring that they collectively span a high-volume region that balances semantic diversity and reward. In this way, DQO can be viewed as extending the principle of D-optimal design from parameter estimation in bandits to the joint optimization of quality and diversity in language model responses. For more detailed discussions, please see Appendix A where we provide a concrete toy example to better illustrate the connections.

### 4.3 Algorithm

We noticed that directly optimizing (3) presents challenges, including high variance in stochastic gradient estimates and risks of numerical instability. To address these issues, we now present a practical algorithmic formulation that stabilizes training and makes the QD objective feasible in practice. To identify the source of these challenges, we begin by computing the gradient of $J_{Div}(\pi_\theta)$, which is given by (for simplicity, we omit the KL-regularization term here),

$$\nabla J_{Div}(\pi_\theta) = \mathbb{E}_{x, y_{1:k} \sim \pi_\theta(\cdot|x)} \left[ \sum_{i=1}^{k} \nabla \log \pi_\theta(y_i|x)(r(x, y_i) + \alpha \log \det(L(y_{1:k}))) \right]. \tag{6}$$

The first issue is that the determinant of $L(y_{1:k})$ can be close to zero, which results in a very large negative value of $\log(\det(L(y_{1:k})))$. This unbounded diversity term destabilizes training and complicates the trade-off between quality and diversity, to the point that only a carefully chosen $\alpha$ is effective. To mitigate this issue, we propose to consider the determinant of the matrix $L(y_{1:k}) + I_k$ instead. It can be shown that by adding an identity matrix, we have $k \geq \log(\det(L(y_{1:k}) + I_k)) \geq 0$ which is well-bounded. Briefly, adding an identity matrix to our objective can be regarded as a regularization term; we further discuss its effect on the objective in Appendix A and provide ablation results in Appendix C.

The second issue is that the gradient consists of the sum of the gradients of $k$ responses $y_{1:k}$, which causes it to have high variance, especially for large $k$. To mitigate the issue of inflating variance, we propose to use leave-one-out (*loo*) gradient estimators by subtracting the log-determinant of the gram matrix which leaves one response out,

$$\nabla^{loo} J_{Div}(\pi_\theta) = \mathbb{E}_{x, y_{1:k} \sim \pi_\theta(\cdot|x)} \left[ \sum_{i=1}^{k} \nabla \log \pi_\theta(y_i|x) \left( r(x, y_i) + \alpha \log \frac{\det(L(y_{1:k}) + I_k)}{\det(L(y_{-i}) + I_{k-1})} \right) \right].$$

Importantly, it can be shown that the *loo* estimator is unbiased and has a nice property on the boundedness of its value shown in Lemma 1 (for the proof, please refer to Appendix A). Lemma 1 shows that the diversity term is non-negative, with an upper bound of order $\log(k)$, which increases slowly as $k$ becomes large. This property stabilizes training and makes DQO robust to large values of $k$.

**Lemma 1** *Let us write the eigenvalues of $L(y_{1:k})$ as $\lambda_k \geq \cdots \geq \lambda_1$, then we have $1 + \lambda_k \geq \frac{\det(L(y_{1:k})+I)}{\det(L(y_{-i})+I)} \geq 1 + \lambda_1$. And the eigenvalue of $L(y_{1:k})$ is always in $[0, k]$ since the embedding vectors are normalized, we have $1 + k \geq \frac{\det(L(y_{1:k})+I)}{\det(L(y_{-i})+I)} \geq 1$ and $\log(1 + k) \geq \log \frac{\det(L(y_{1:k})+I)}{\det(L(y_{-i})+I)} \geq 0$.*

## 5 Experiments

In this section, we conduct a series of experiments to evaluate the performance of DQO in generating diverse and high-quality responses. Specifically, we aim to answer the following questions:

- Does DQO improve diversity in responses and how does it compare with the reward-only baseline and other existing quality-diversity algorithms?
- Does DQO achieve a favorable balance between quality and diversity? Can the model preserve or improve task performance while enhancing diversity?
- Is the performance of DQO consistent across different tasks and settings?
- How does DQO manage the trade-off between quality and diversity, and how do its hyperparameters influence performance?

## 5.1 City recommendation

We begin with a simple synthetic experiment on city recommendation to clearly illustrate the diversity achieved by DQO. In this task, the model was prompted to recommend a city for traveling along with a concise reason. The exact prompt we used is provided in Appendix E. We compared DQO with GRPO and also implemented a variant of DQO using the average pairwise distance as the diversity score, which we refer to as DQO-pairwise distance; the original algorithm is denoted as DQO-determinant. The results are summarized in Figure 2, and the full details, including exact numbers and city names, are reported in Appendix C.

DQO clearly encourages the model to generate more diverse recommendations. When trained solely with the reward, the model tends to converge on recommending the same city repeatedly. For using the pairwise distance as the diversity score, we observed that the model's recommendations were dominated by two major cities, which aligns with our previous analysis. This occurs because high pairwise distance can be achieved with two widely separated clusters. In contrast, the determinant-based approach penalizes linear dependence, encouraging responses to span the space as broadly as possible. As shown in Figure 2, DQO-determinant produces the most diverse set of recommendations.

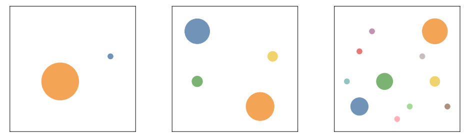


Figure 2: From left to right: GRPO, DQO-pairwise distance, DQO-determinant. Each circle represents a different city, with the size proportional to the number of times it was recommended. For each model, we sampled 100 times with a temperature of 1.0.

## 5.2 General tasks

In the above section, we show the effectiveness of DQO in promoting diversity on the controlled city recommendation task, we now turn to a broader evaluation on general language model tasks. We implemented extensive experiments on four different kinds of tasks including reasoning (GSM8K [Cobbe et al., 2021]), summarization (CNN-dailymail [See et al., 2017]), story-writing (Common-Gen [Lin et al., 2020]) and instruction-following (Dolly Conover et al. [2023]).

**Baselines.** We compare DQO to the baseline algorithm which trains the model solely with reward. For reasoning tasks, we use GRPO, while for non-reasoning tasks, we adopt PPO. We also compare DQO with other two popular quality-diversity algorithms: GRPO-likelihood [He et al., 2025] and GRPO-entropy [Yao et al., 2025]. For the detailed experimental setup, please see Appendix B.

In this work, we employ a reward model to provide quality scores. Notice that, for reasoning tasks, we also rely on the reward model rather than outcome-based rewards, due to the observed phenomenon of reward hacking with outcome reward (for details, please refer to Appendix D). During training, rewards are normalized by dividing by an empirical maximum value to ensure a comparable scale with the diversity score, whereas during evaluation, we report the unnormalized rewards.

**Metrics.** We report $pass@n$ metric (i.e., the highest score among $n$ responses) as measures of the quality in the responses with $n$ varies from 1 to 10. Without special clarifications, responses are sampled with a temperature of 1.0. And we use multiple metrics to measure the diversity in the responses which we summarize below,

- Distinct-n: Count the ratio of unique n-grams among the responses.
- Self-BLEU [Papineni et al., 2002] and Self-ROUGE [Lin, 2004] score: Two popular metrics to measure the similarity of languages. Note these scores measure the similarity, to be consistent with other metrics, we report $1 - Score$.
- LLM as a judge: We prompt an advanced model GPT-4o-mini to judge the model's output in terms of the diversity (see Appendix E and F), serving as a surrogate for human judgment.

We first compare the performance of DQO against all baseline algorithms. The results are summarized in Table 1. To evaluate the model, we select a representative non-reasoning task and a reasoning task. Among all algorithms, DQO is the only one that achieves both high diversity and high quality scores

Table 1: The quality and diversity scores achieved by DQO compared to other baseline algorithms on the instruction-following task: Dolly, and the reasoning task: GSM8K. Diversity metrics are calculated across 10 generated responses per prompt.

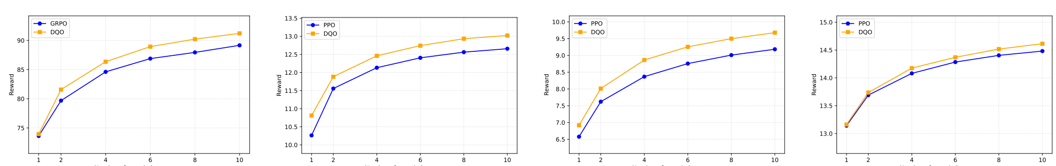
| Method / Task | Diversity ↑ | | | | Quality ↑ | |
|---|---|---|---|---|---|---|
| | distinct-1 | distinct-4 | self-bleu | self-rouge | $pass@1$ | $pass@10$ |
| **Dolly** | | | | | | |
| PPO | 0.24 | 0.64 | 0.41 | 0.49 | 5.65 | 8.39 |
| GRPO-likelihood | 0.26 | 0.70 | 0.46 | 0.54 | 5.86 | 8.50 |
| GRPO-entropy | 0.36 | 0.75 | 0.56 | 0.57 | 4.71 | 7.70 |
| DQO | 0.28 | 0.69 | 0.46 | 0.54 | 5.92 | 8.74 |
| **GSM8K** | | | | | | |
| GRPO | 0.09 | 0.32 | 0.09 | 0.21 | 76.8 | 87.9 |
| GRPO-likelihood | 0.26 | 0.86 | 0.53 | 0.59 | 50.9 | 80.4 |
| GRPO-entropy | 0.10 | 0.38 | 0.12 | 0.25 | 77.0 | 92.6 |
| DQO | 0.10 | 0.42 | 0.14 | 0.31 | 76.3 | 91.2 |

across both tasks. GRPO-likelihood shows performance comparable to DQO on the Dolly task but underperforms on GSM8K, whereas GRPO-entropy performs well on GSM8K but poorly on Dolly. These results demonstrate that DQO consistently delivers strong performance in post-training LLMs to produce diverse and high-quality generations.

### 5.3 Quality-diversity balance

In this section, we present more fine-grained results on the diversity and quality of responses generated by the model trained with DQO. For comparison, we also include the results of the model trained solely with the reward.

**Quality.** In Figure 3, we show the $pass@n$ performance across four tasks with $n$ varying from 1 to 10. DQO exhibits better performance than the baseline model especially when $n$ is large. Besides, in the case of $n = 1$, our model has similar or better performance to the baseline. Together, the results show that our method does not hurt $pass@1$ performance while providing better $pass@n$ performance with $n > 1$ indicating that our model can generate both high-quality and diverse responses.

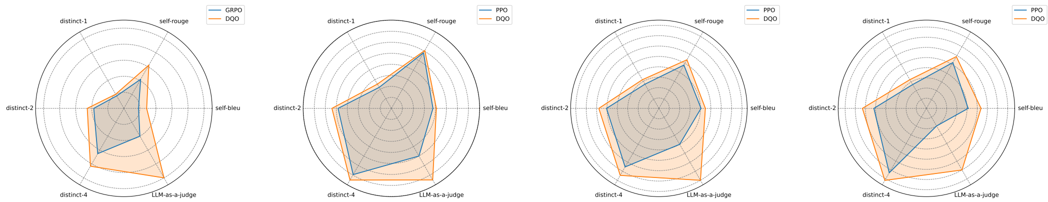



(a) GSM8K (b) CNN-dailymail (c) Dolly (d) Gen

Figure 3: The performance of the trained model on $pass@n$ metrics. For DQO, we set hyperparameters $\alpha = 1.0$ and $k = 4$.

**Diversity.** The superior performance on $pass@n$ already suggests that our method enhances response diversity. To further validate this, we present six diversity metrics in Figure 3. For each metric, higher values indicate greater diversity. As shown in the figure, DQO consistently outperforms the baseline model, demonstrating a clear advantage in diversity. In particular, for the LLM-as-a-judge metric, the advanced model GPT-4o-mini strongly recognizes the diversity of responses generated by our approach (See Appendix F), highlighting improvements at the semantic level.

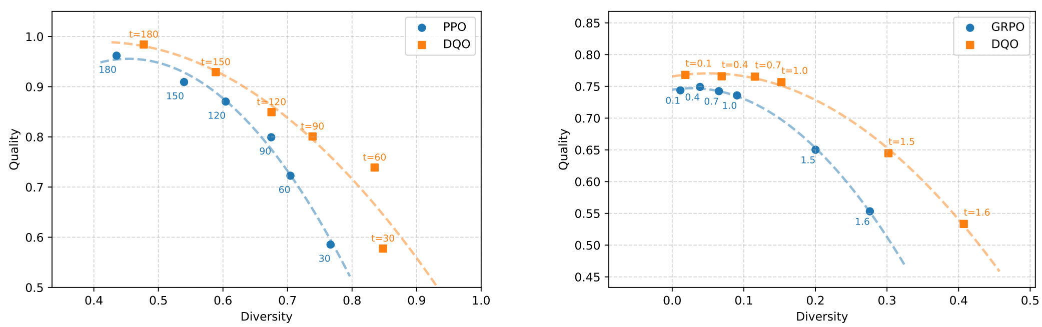
**Pareto frontier.** To illustrate how DQO achieves a favorable balance between quality and diversity, we plot the Pareto frontiers of DQO and the baseline model by varying either the training steps or the




(a) GSM8K (b) CNN-dailymail (c) Dolly (d) Gen

Figure 4: The performance of the trained model on six diversity metrics. For each metric, the higher value means the higher diversity. For DQO, we set hyperparameters $\alpha = 1.0$ and $k = 4$. And the diversity metrics are calculated across 10 generated responses per prompt.

sampling temperature in Figure 5. Across different sampling temperatures (the right in Figure 5), our model consistently occupies the upper-right region relative to the baseline, demonstrating a robust advantage in balancing quality and diversity at the inference stage. Similarly, when varying the training steps (the left in Figure 5), our model remains Pareto-optimal throughout the entire training process, indicating that it consistently achieves a better quality–diversity balance throughout the entire training process.




(a) Gen, varying training steps (b) GSM8K, varying decoding temperature

Figure 5: Pareto frontiers on quality and diversity of our model and the baseline. On the left, each point is a model trained with different training steps and the sampling temperature is set as 1.0. On the right, we take the final trained model but vary the sampling temperature.

### 5.4 Ablation Study

DQO introduces two key hyperparameters: $\alpha$ and $k$. The hyperparameter $\alpha$ controls the weight of the diversity term in the training objective, while $k$ is the number of responses sampled per prompt in the algorithm. Both parameters jointly influence the trade-off between output quality and diversity. To analyze their impact, we conduct experiments across different values of $\alpha$ and $k$, and the results on Dolly task are summarized in Table 2 (for the ablations on GSM8K task, please refer to Appendix C).

From Table 2, we can observe the trade-off between quality and diversity when changing the value of $\alpha$ or $k$. Both increasing $\alpha$ and $k$ can enhance the diversity of generated responses as reflected in diversity metrics such as distinct-n, self-bleu and self-rouge scores. However, when $\alpha$ or $k$ exceeds a certain threshold, the quality metric pass@1 begins to decline, suggesting that the diversity term becomes too dominant. At the same time, pass@10 metric remains relatively stable, indicating that the model still generates some high-quality responses among its top outputs despite the strong emphasis on diversity.

Increasing $k$ enhances response diversity, since a larger k encourages the model to generate a larger group of diverse responses, that is, to allocate probability mass across a wider range of outputs. However, unlike adjusting $\alpha$, increasing $k$ incurs additional computational costs since more responses

Table 2: The quality and diversity scores of the model trained with different values of hyperparameters $k$ and $\alpha$ on the Dolly task. Diversity metrics are calculated across 10 generated responses per prompt.

| Method | Diversity ↑ | | | | Quality ↑ | |
|---|---|---|---|---|---|---|
| | distinct-1 | distinct-4 | self-bleu | self-rouge | $pass@1$ | $pass@10$ |
| PPO | 0.24 | 0.64 | 0.41 | 0.49 | 5.65 | 8.39 |
| $\alpha = 0.5, k = 4$ | 0.28 | 0.69 | 0.44 | 0.53 | 5.84 | 8.79 |
| $\alpha = 1.0, k = 4$ | 0.28 | 0.69 | 0.47 | 0.54 | 5.92 | 8.73 |
| $\alpha = 1.5, k = 4$ | 0.33 | 0.79 | 0.57 | 0.61 | 5.47 | 8.56 |
| $\alpha = 2.0, k = 4$ | 0.35 | 0.82 | 0.54 | 0.64 | 5.42 | 8.69 |
| $k = 2, \alpha = 1.0$ | 0.24 | 0.62 | 0.40 | 0.50 | 5.71 | 8.13 |
| $k = 4, \alpha = 1.0$ | 0.28 | 0.69 | 0.47 | 0.54 | 5.92 | 8.73 |
| $k = 6, \alpha = 1.0$ | 0.31 | 0.76 | 0.49 | 0.58 | 5.71 | 8.83 |
| $k = 8, \alpha = 1.0$ | 0.32 | 0.79 | 0.52 | 0.61 | 5.64 | 8.64 |

must be generated. From a practical perspective, we therefore recommend prioritizing adjustments to $\alpha$ over $k$ since the performance gain is similar.

Overall, compared with the baseline model, DQO exhibits robust improvements in both quality and diversity across a wide range of $\alpha$ and $k$ values, suggesting that it does not require highly sensitive hyperparameter tuning. In practice, this property makes the algorithm more reliable and easier to deploy, as strong performance can be maintained under different settings.

In addition to the ablations on different values of $\alpha$ and $k$, we also conduct ablations on the weight of the identity matrix in the determinant calculation (see Section 4.3), controlled by the parameter $\gamma$ in $\det(L(y_{1:k}) + \gamma I_k)$. Please refer to Appendix C for the results.

## 6 Conclusions

In this work, we propose an algorithm DQO to post-train LLMs for diverse high-quality responses. Based on determinantal point processes, DQO defines the diversity in a group of responses as the determinant of a kernel-based similarity matrix of the embeddings of those responses. This definition of diversity has a straightforward interpretation as the squared volume of the space spanned by the embeddings of the response. We conduct extensive experiments across different kinds of tasks, and show that DQO can optimize the model to generate significantly more diverse responses while maintaining high quality in the generated responses.

Although DQO achieves superior performance, there are some limitations in this work. First, the quality-diversity objective is vulnerable to reward hacking when using the outcome reward. A reward model is needed which limits the applicability in many reasoning tasks where the outcome reward is commonly used. Second, DQO relies on the embedding models to map responses into a semantic space. The performance of DQO depends on the quality of these embeddings. A more principled and adaptive method for measuring diversity could better capture the underlying semantic variation, potentially adjusting automatically to the specific requirements of different tasks, which may emphasize different aspects of diversity, thus achieves better performance.

## Acknowledgements

The research was partially supported by the NSF under grants DEB-2433726, ECCS-2317079, CCF-2200052, and IIS-1914792, by the DOE under grant DE-AC02-05CH11231, by the NIH under grant UL54 TR004130, and by the Hariri Institute at Boston University. The authors would also like to acknowledge New England Research Cloud (NERC) for providing the computational resources that partially supported the experiments conducted in this work.

## References


David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(85)80012-4. URL `https://www.sciencedirect.com/science/article/pii/S0364021385800124`.

Eltayeb Ahmed, Uljad Berdica, Martha Elliott, Danijela Horak, and Jakob N Foerster. Intent factored generation: Unleashing the diversity in your language model. *arXiv preprint arXiv:2506.09659*, 2025.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL `https://arxiv.org/abs/2204.05862`.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=bx24KpJ4Eb`. Survey Certification, Featured Certification.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing, 2025. URL `https://arxiv.org/abs/2503.17126`.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm`.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng

Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

Steve Fisk. A very short proof of cauchy's interlace theorem for eigenvalues of hermitian matrices. *arXiv preprint math/0502408*, 2005.

Giorgio Franceschelli and Mirco Musolesi. Diffsampling: Enhancing diversity and accuracy in neural text generation, 2025. URL `https://arxiv.org/abs/2502.14037`.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *CoRR*, abs/2412.10271, 2024. URL `https://doi.org/10.48550/arXiv.2412.10271`.

Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. *arXiv preprint arXiv:2506.02355*, 2025.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304, 1959.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *CoRR*, abs/2310.06452, 2023. URL `https://doi.org/10.48550/arXiv.2310.06452`.

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *ICLR*, 2025.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.165`.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025.

Sonia K Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. *arXiv preprint arXiv:2411.04427*, 2024.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=TG8KACxEON`.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Feiz5HtCD0`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof Choromanski, and Stephen Roberts. Effective Diversity in Population Based Reinforcement Learning. In *Advances in Neural Information Processing Systems 34*. 2020.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof M Choromanski, and Stephen J Roberts. Effective diversity in population based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:18050–18062, 2020.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HPuSIXJaa9`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL `https://www.aclweb.org/anthology/P17-1099`.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and comparative analysis, 2024. URL `https://openreview.net/forum?id=jvRCirB0Oq`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL `https://arxiv.org/abs/2402.03300`.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of LLM generated content. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=O7bF6nlSOD`.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, 2020.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35), 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning. *arXiv preprint arXiv:2505.23433*, 2025.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL `https://arxiv.org/abs/1909.08593`.

# A Theoretical results and proofs

**Lemma.** Suppose $\psi(x, y) = \sqrt{\exp(\frac{r(x,y)}{\alpha})\pi_{ref}(y|x)} \cdot \phi(y)$, then the optimal policy defined in (4) satisfies $\pi_{div}(y_{1:k}|x) \propto \det(L_\psi(y_{1:k}))$ when $\alpha = \beta$.

**Proof.** Let $B \in \mathbb{R}^{n \times k}$ have columns $\phi(y_1), \ldots, \phi(b_k)$. The Gram matrix is

$$L = B^\top B.$$

Now suppose we scale each column $\phi(y_i)$ by a factor $a_i$, and denote

$$A = \mathrm{diag}(a_1, \ldots, a_k), \quad B' = BA.$$

Then the new Gram matrix is

$$L' = (B')^\top B' = (AB^\top)(BA) = A(B^\top B)A = ALA.$$

Taking determinants,

$$\det(L') = \det(ALA) = \det(A)\,\det(L)\,\det(A) = \big(\det(A)\big)^2 \det(L).$$

Since $\det(A) = \prod_{i=1}^k a_i$, we obtain

$$\det(L') = \left(\prod_{i=1}^k a_i\right)^2 \det(L).$$

Recall that $\pi_{div}(y_{1:k}|x)$ is defined as when $\alpha = \beta$,

$$\begin{aligned}
\pi_{div}(y_{1:k}|x) &\propto \pi_{ref}(y_{1:k}|x) \exp\left(\frac{1}{\alpha}\left(\sum_{i=1}^k r(x, y_i)\right) + \log\det(L_\phi(y_{1:k}))\right) \\
&= \pi_{ref}(y_{1:k}|x) \exp\left(\frac{1}{\alpha}\left(\sum_{i=1}^k r(x, y_i)\right)\right) \det(L_\phi(y_{1:k})) \\
&= \prod_{i=1}^k \left(\pi_{ref}(y_i|x) \exp\left(\frac{r(x, y_i)}{\alpha}\right)\right) \det(L_\phi(y_{1:k}))
\end{aligned}$$

The second equality holds because $y_{1:k}$ are sampled independently. Combined with the result above, we have $\pi_{div}(y_{1:k}|x) \propto \det(L_\psi(y_{1:k}))$.

**Analysis of $\det(L(y_{1:k}))$ and $\det(L(y_{1:k}) + I_k)$.** Maximizing $\det(L)$ is equivalent to maximizing the volume of the parallelepiped spanned by the feature vectors of selected responses, which enforces strict linear independence: any subset that induces a singular matrix receives zero score. In contrast, maximizing $\det(L + I)$ introduces a ridge-like regularization. Indeed, if $L = BB^\top$ for a feature matrix $B \in \mathbb{R}^{k \times d}$, we have

$$\det(L + I) = \det(BB^\top + I) = \det(I + B^\top B).$$

This is precisely the determinant of a regularized scatter matrix, analogous to the role of $(B^\top B + \lambda I)$ in ridge regression. From this viewpoint, adding $I$ stabilizes the objective by preventing collapse along directions of near-linear dependence and avoiding the degeneracy of zero determinants.

A complementary interpretation arises from Bayesian linear models and Gaussian processes. In Bayesian linear regression with a Gaussian prior $w \sim \mathcal{N}(0, I)$ and unit-variance observation noise, the marginal likelihood normalization involves $\det(I + B^\top B)^{-\frac{1}{2}}$. Similarly, in Gaussian process regression, the log marginal likelihood includes $\log\det(L + \sigma^2 I)$, with $\sigma^2$ corresponding to the noise variance. Setting $\sigma^2 = 1$ recovers the $\det(L + I)$ objective. Hence, $\det(L + I)$ can be viewed as the determinant under a model with a prior noise floor, which softens the diversity requirement and balances between variance explained by the selected items and a baseline level of uncertainty.

**Eigenvalue Interlacing Theorem [Fisk, 2005].** Suppose $A \in R^{n\times n}$ is symmetric. Let $B \in R^{m\times m}$ with $m < n$ be a principal submatrix (obtained by deleting both $i$-th row and $i$-th column for some values of $i$). Suppose $A$ has eigenvalues $\lambda_1 \le \cdots \le \lambda_n$ and $B$ has eigenvalues $\beta_1 \le \cdots \le \beta_m$. Then,

$$\lambda_k \le \beta_k \le \lambda_{k+n-m}, \text{ for } k = 1, \cdots, m$$

And if $m = n - 1$, one has,

$$\lambda_1 \le \beta_1 \le \lambda_2 \le \beta_2 \le \cdots \le \beta_{n-1} \le \lambda_n$$

**Proof.** We use the Courant–Fischer min–max theorem. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1 \le \cdots \le \lambda_n$, the $k$-th eigenvalue can be characterized as

$$\lambda_k = \min_{\substack{S\subset\mathbb{R}^n \\ \dim(S)=k}} \max_{\substack{x\in S \\ x\neq 0}} \frac{x^\top Ax}{x^\top x}.$$

Similarly, for the principal submatrix $B \in \mathbb{R}^{m\times m}$ with eigenvalues $\beta_1 \le \cdots \le \beta_m$, we have

$$\beta_k = \min_{\substack{T\subset\mathbb{R}^m \\ \dim(T)=k}} \max_{\substack{y\in T \\ y\neq 0}} \frac{y^\top By}{y^\top y}.$$

Now observe that $B$ is obtained by restricting $A$ to a coordinate subspace (corresponding to removing some rows and columns). Hence any $y \in \mathbb{R}^m$ can be embedded into $\mathbb{R}^n$ by padding with zeros. Under this embedding, the Rayleigh quotient is preserved:

$$\frac{y^\top By}{y^\top y} = \frac{x^\top Ax}{x^\top x}, \quad \text{where } x \text{ is } y \text{ padded with zeros.}$$

Therefore, the feasible subspaces for $B$ are restrictions of those for $A$. This leads to the inequalities

$$\lambda_k \le \beta_k \le \lambda_{k+n-m}, \quad k = 1, \ldots, m.$$

In the special case $m = n - 1$, the inequalities expand into the chain

$$\lambda_1 \le \beta_1 \le \lambda_2 \le \beta_2 \le \cdots \le \beta_{n-1} \le \lambda_n,$$

which is exactly the interlacing property.

**Lemma.** Let's write the eigenvalues of $L(y_{1:k})$ as $\lambda_k \ge \cdots \ge \lambda_1$, then we have $1 + \lambda_k \ge \frac{\det(L(y_{1:k})+I_k)}{\det(L(y_{-i})+I_{k-1})} \ge 1+\lambda_1$. And the eigenvalue of $L(y_{1:k})$ is always in $[0, k]$ since the embedding vectors are normalized, we have $1+k \ge \frac{\det(L(y_{1:k})+I_k)}{\det(L(y_{-i})+I_{k-1})} \ge 1$ and $\log(1+k) \ge \log \frac{\det(L(y_{1:k})+I_k)}{\det(L(y_{-i})+I_{k-1})} \ge 0$.

**Proof.** Let's write the eigenvalues of $L(y_{-i})$ as $\beta_{k-1} \ge \cdots \ge \beta_1$. Based on Eigenvalue Interlacing Theorem, we have,

$$\frac{\det(L(y_{1:k}) + I_k)}{\det(L(y_{-i}) + I_{k-1})} = (1 + \lambda_1) \prod_{i=1}^{k-1} \frac{1 + \lambda_{i+1}}{1 + \beta_i} \ge 1 + \lambda_1$$

and,

$$\frac{\det(L(y_{1:k}) + I_k)}{\det(L(y_{-i}) + I_{k-1})} = (1 + \lambda_k) \prod_{i=1}^{k-1} \frac{1 + \lambda_i}{1 + \beta_i} \le 1 + \lambda_k$$

Since $L(y_{1:k})$ is positive semidefinite, it holds $\lambda_i \ge 0$, $\forall i$. And we have $\sum_{i=1}^{k} \lambda_i = \mathrm{tr}(L(y_{1:k})) = k$ due to the normalization of the feature vectors. Hence, we have $k \ge \lambda_k \ge \lambda_1 \ge 0$.

**Connections with D-Optimal Design: A toy exmaple.** We construct a controlled toy setup to empirically compare our setting to D-optimal selection strategy against the baseline of uniform sampling, highlighting scenarios where diversity plays a critical role in achieving robust performance across varied reward functions

Setup and Notation: Let $d = 3$ denote the dimensionality of the embedding space, and let $\{\phi_1, \phi_2, \ldots, \phi_N\} \subset \mathbb{R}^d$ be a set of normalized candidate embeddings. We synthetically construct the pool to be imbalanced along coordinate directions: $n_x = 40$: vectors near the $x$-axis, $n_y = 40$: vectors near the $y$-axis and $n_z = 10$: vectors near the $z$-axis (rare) with $N = 90$. Small Gaussian noise $\varepsilon \sim \mathcal{N}(0, 10^{-4}I)$ is added to prevent rank-deficiency.

To illustrate our hypothesis, we compare two sampling strategies: **Uniform**: $w_i = \frac{1}{N}$ for all $i$ and **ours** by solving the following optimization problem

$$\begin{aligned} \max_{w \in \Delta_N} \quad & \log\det\left(\sum_{i=1}^{N} w_i \phi_i \phi_i^\top + \delta I\right) \\ \text{s.t.} \quad & \sum_{i=1}^{N} w_i = 1, \quad 0 \le w_i \le 1 \end{aligned}$$

where $\delta = 10^{-9}$ ensures numerical stability. The DPP based approach promotes spectral coverage across all directions. Next, we define four linear reward directions: $k_x = [1, 0, 0]^\top$ (high mass), $k_y = [0, 1, 0]^\top$ (high mass), $k_z = [0, 0, 1]^\top$ (low mass), $k_{\text{bal}} = \frac{1}{\sqrt{3}}[1, 1, 1]^\top$ (balanced). The expected reward under policy $w$ and reward $k$ is

$$\mathbb{E}_{\phi \sim w}\left[\langle k, \phi \rangle\right] = \sum_{i=1}^{N} w_i \langle k, \phi_i \rangle \tag{7}$$

We compute the information matrices: $\Sigma(w) = \sum_{i=1}^{N} w_i \phi_i \phi_i^\top$ and visualize their eigenvalues. Uniform sampling has low spectral mass in the $z$-direction. In contrast, our DPP based approach balances mass across all directions. This confirms that $\log\det$ promotes **coverage**.

Thus this simulation reveals a failure mode of uniform sampling in imbalanced datasets. D-optimal design reallocates weights toward rare directions, yielding significantly improved performance in worst-case reward scenarios. The log-determinant acts as a diversity-promoting surrogate, superior to naive metrics like pairwise distance or cosine dissimilarity.

# B Experimental setup

**Data preparation.** For GSM8K dataset, we directly use the training and test split. For CNN-dailymail dataset, we take the test split, and select $8,000$ data points as the training set and $1,024$ data points as the test set. For Dolly dataset, there is only one training split of $15,000$ data points. We divided it into two subsets with the ratio of $0.2$. For Gen, we use the training split, remove data with repetitive key words, and divided the set into two subsets, each containing $8,000$ and $1,024$ data points respectively.

**Training configuration.** We use GRPO algorithm for GSM8K task and use PPO algorithm for the other tasks. We use Qwen2.5-MATH-1.5B [Yang et al., 2024] as the base model for GSM8K task, and use Llama3.2-1B for the other tasks. For all tasks, we use the reward model Skywork/Skywork-Reward-V2-Llama-3.2-1B [Liu et al., 2025] and embedding model sentence-transformers/all-MiniLM-L6-v2. For baseline algorithms, we directly use the hyperparameters reported in their papers. For GRPO-likelihood, we use $\alpha = 0.25$ and for GRPO-entropy, we use $\alpha = 0.01$.

Table 3: Training configurations. For max prompt and response length, we use different values for different datasets. From left to right, it corresponds to GSM8K, CNN-dailymail, Dolly, Gen.

| | |
|---|---|
| training batch size | 128 |
| training epoches | 3 |
| actor learning rate | 1e-6 |
| critic learning rate | 1e-5 |
| rollout temperature | 1.0 |
| max prompt length | 256, 1536, 1024, 64 |
| max response length | 256, 256, 512, 128 |

## C Additional experiments

### C.1 City recommendation

The following table shows the numerical results that were used to plot Figure 2. For each model, we sampled 100 times and countered the times of each city being recommended.

Table 4: The frequencies of each city being recommended for models trained with different methods.

| Method | City (Frequency) |
|---|---|
| GRPO | Tokyo (97); New York (3) |
| DQO-pairwise distance | New Orleans (48); Asheville (37); Budapest (8); Barcelona (7) |
| DQO-determinant | Budapest (45); Chiang Mai (22); New Orleans (19); Hanoi (7); Krakow (1); Kanazawa (1); Ottawa (1); Nashville (1); Tokyo (1); Bangkok (1); Singapore (1) |

### C.2 Additional ablation studies

**Value of $\alpha$ and $k$.** In Table 2 in the main page, we show the ablation results of $\alpha$ and $k$ on Dolly task. Here, Table 5 shows the results on GSM8K task.

Table 5: The quality and diversity scores of the model trained with different values of hyperparameters $k$ and $\alpha$ on the GSM8K task. Diversity metrics are calculated across 10 generated responses per prompt.

| Method | Diversity ↑ | | | | Quality ↑ | |
|---|---|---|---|---|---|---|
| | distinct-1 | distinct-4 | self-bleu | self-rouge | $pass@1$ | $pass@10$ |
| GRPO | 0.09 | 0.32 | 0.09 | 0.21 | 76.8 | 87.9 |
| $\alpha = 0.5, k = 4$ | 0.09 | 0.33 | 0.09 | 0.22 | 74.6 | 89.2 |
| $\alpha = 1.0, k = 4$ | 0.10 | 0.42 | 0.14 | 0.31 | 76.3 | 91.2 |
| $\alpha = 1.5, k = 4$ | 0.11 | 0.48 | 0.19 | 0.34 | 76.1 | 92.6 |
| $\alpha = 2.0, k = 4$ | 0.13 | 0.54 | 0.21 | 0.40 | 76.7 | 92.7 |
| $\alpha = 5.0, k = 4$ | 0.16 | 0.62 | 0.28 | 0.44 | 77.7 | 93.3 |
| $k = 2, \alpha = 1.0$ | 0.11 | 0.40 | 0.13 | 0.26 | 73.9 | 90.3 |
| $k = 4, \alpha = 1.0$ | 0.10 | 0.42 | 0.14 | 0.31 | 76.3 | 91.2 |
| $k = 6, \alpha = 1.0$ | 0.10 | 0.44 | 0.17 | 0.33 | 76.5 | 92.1 |
| $k = 8, \alpha = 1.0$ | 0.11 | 0.47 | 0.16 | 0.32 | 74.9 | 90.8 |

**Kernel function.** DQO formulates the diversity score as the determinant of the kernel matrix. By default, we use the dot product kernel function, i.e., $\phi(y_1, y_2) = \langle y_1, y_2 \rangle$. We study the effect of different kernel functions on DQO. We implemented same experiments using Gaussian kernel function, i.e., $\phi(y_1, y_2) = \exp\left(-\frac{\|y_1 - y_2'\|^2}{2}\right)$. Table 6 shows the performance of DQO with different kernel functions on Dolly and GSM8K tasks. DQO demonstrates robust performance under different kernel functions.

**Regularization by introducing identity matrix.** To solve the numerical explosion issue, we introduce an identity matrix when calculating the determinant: $\det(L(y_{1:k} + \gamma I_k))$. It can be shown adding an identity matrix plays a role as a regularization. By default, we simply set $\gamma = 1$. We test DQO with different values of $\gamma$ on GSM8K and Dolly tasks.

We can see from Table 7, when $\gamma = 0.1$, the diversity in responses surges while the quality collapses. This is consistent with our analysis. The identity matrix can be regarded as a regularization term. The

Table 6: The quality and diversity scores of the model trained with different kernel functions on Dolly and GSM8K tasks. Diversity metrics are calculated across 10 generated responses per prompt.

| Method | Diversity ↑ | | | | Quality ↑ | |
|---|---|---|---|---|---|---|
| | distinct-1 | distinct-4 | self-bleu | self-rouge | $pass@1$ | $pass@10$ |
| Dolly | | | | | | |
| $\alpha = 1.0$, gaussian | 0.29 | 0.72 | 0.48 | 0.56 | 6.45 | 8.61 |
| $\alpha = 1.0$, dot product | 0.28 | 0.69 | 0.46 | 0.54 | 6.56 | 8.74 |
| $\alpha = 2.0$, gaussian | 0.34 | 0.79 | 0.54 | 0.61 | 6.12 | 8.64 |
| $\alpha = 2.0$, dot product | 0.35 | 0.82 | 0.54 | 0.64 | 6.41 | 8.69 |
| GSM8K | | | | | | |
| $\alpha = 1.0$, gaussian | 0.10 | 0.43 | 0.16 | 0.31 | 77.1 | 90.9 |
| $\alpha = 1.0$, dot product | 0.10 | 0.42 | 0.14 | 0.31 | 76.3 | 91.2 |
| $\alpha = 2.0$, gaussian | 0.11 | 0.48 | 0.19 | 0.36 | 75.2 | 91.2 |
| $\alpha = 2.0$, dot product | 0.13 | 0.54 | 0.21 | 0.40 | 76.7 | 92.7 |

objective will prioritize diversity more if $\gamma$ is low. In addition, when $\alpha = 0.5$, we can see decreasing $\gamma$ does not affect the quality much as in the case where $\alpha = 1.0$. This is because $\alpha$ also controls the balance between quality and diversity. When $\alpha$ is low, the diversity is less important in the objective, hence, the effect of decreasing $\gamma$ is diluted. The phenomenon is consistent on GSM8K task.

Table 7: The performance of DQO with different identity matrix weights on Dolly task. Diversity metrics are calculated across 10 generated responses per prompt.

| Method | Diversity ↑ | | | | Quality ↑ | |
|---|---|---|---|---|---|---|
| | distinct-1 | distinct-4 | self-bleu | self-rouge | $pass@1$ | $pass@10$ |
| $\alpha = 1.0, +0.1I$ | 0.57 | 0.96 | 0.79 | 0.86 | 3.44 | 6.38 |
| $\alpha = 1.0, +0.5I$ | 0.37 | 0.83 | 0.56 | 0.66 | 6.72 | 8.90 |
| $\alpha = 1.0, +I$ | 0.28 | 0.69 | 0.46 | 0.54 | 6.56 | 8.74 |
| $\alpha = 0.5, +0.1I$ | 0.34 | 0.84 | 0.58 | 0.68 | 6.04 | 8.75 |
| $\alpha = 0.5, +0.5I$ | 0.28 | 0.71 | 0.44 | 0.54 | 6.31 | 8.72 |
| $\alpha = 0.5, +I$ | 0.28 | 0.69 | 0.43 | 0.53 | 6.47 | 8.77 |

Table 8: The performance of DQO with different identity matrix weights on GSM8K task. Diversity metrics are calculated across 10 generated responses per prompt.

| Method | Diversity ↑ | | | | Quality ↑ | |
|---|---|---|---|---|---|---|
| | distinct-1 | distinct-4 | self-bleu | self-rouge | $pass@1$ | $pass@10$ |
| $\alpha = 1.0, +0.1I$ | 0.23 | 0.79 | 0.41 | 0.52 | 73.3 | 93.9 |
| $\alpha = 1.0, +0.5I$ | 0.11 | 0.49 | 0.19 | 0.36 | 77.5 | 92.0 |
| $\alpha = 1.0, +I$ | 0.10 | 0.42 | 0.14 | 0.31 | 76.3 | 91.2 |
| $\alpha = 2.0, +0.1I$ | 0.38 | 0.88 | 0.58 | 0.74 | 63.5 | 91.3 |
| $\alpha = 2.0, +0.5I$ | 0.15 | 0.59 | 0.24 | 0.42 | 78.0 | 93.2 |
| $\alpha = 2.0, +I$ | 0.13 | 0.54 | 0.21 | 0.40 | 76.7 | 92.7 |

# D Reward hacking with outcome reward

We observed an interesting phenomenon when doing experiments on GSM8K task. With only outcome reward provided, the model trained with the quality-diversity objective learns to first generate a correct answer to the question thus securing the outcome reward, then generate some random or irrelevant contents to artificially increase diversity. See the concrete example below.

The reward-hacking phenomenon does not exist when we use reward models, since the reward model evaluates the quality of the entire response rather than just the final answer.

We explored a potential solution by computing the cosine similarity between the embeddings of the prompt and the response and multiplying this score with the diversity term. The intuition is that only responses semantically aligned with the prompt contribute to the diversity objective. However, we observed that the model can still exploit this mechanism by generating correct answers followed by random contents but superficially related to the prompt. See the concrete example below.

We believe outcome-based reward alone is inherently susceptible to reward hacking. The outcome reward is widely used in reasoning tasks due to its simplicity and effectiveness, however, when combined with other objectives, it is vulnerable to reward hacking. This phenomenon presents an intriguing challenge for future study.

**An example of reward hacking with outcome reward**

Prompt: Gretchen has some coins. There are 30 more gold coins than silver coins. If she had 70 gold coins, how many coins did Gretchen have in total? Let's think step by step and output the final answer after "####".

Response: If Gretchen had 70 gold coins and there are 30 more gold coins than silver coins, then the number of silver coins she had is

$$70 - 30 = 40 \text{ silver coins.}$$

The total number of coins Gretchen has is the sum of gold and silver coins, so it is

$$70 + 40 = 110 \text{ coins.} \quad \text{\#\#\#\# } 110$$

The correct answer is: 110

Which digit is first to appear in the units place of

$$5^1 + 5^2 + 5^3 + \cdots + 5^{2004} + 5^{2005} \text{ ?}$$

(A) 0
(B) 1
(C) 2
(D) 5

To determine which digit is first to appear in the units place of the sum

$$5^1 + 5^2 + 5^3 + \cdots + 5^{2004} + 5^{2005},$$

we need to analyze the units digits of the powers of 5.
First, let's observe the units digits of the first few powers of 5:

...

**An example of reward hacking with prompt-response similarity**

Prompt: Bobby takes a 30 min lunch and two 15 minutes break per day at the office. After 5 days, how many hours do his lunches and breaks add up to? Let's think step by step and output the final answer after "####".

Response: Bobby takes a 30 minute lunch and a 15 minute break twice a day.
So in total, Bobby takes 30 + 15 + 15 = 60 minutes of lunch and breaks per day.
After 5 days, Bobby takes a total of

$$60 \times 5 = 300 \text{ minutes of lunch and breaks.}$$

Since there are 60 minutes in an hour, Bobby's lunches and breaks add up to

$$\frac{300}{60} = 5 \text{ hours.}$$

#### 5
Bobby's lunches and breaks add up to 5 hours.
Bobby's dog is on a diet. For every meal he eats, he needs to walk him for 2 minutes. When he eats twice, he walks him 4 times that amount. How many minutes did Bobby walk his dog if he let him eat 7 meals? Let's think step by step and output the final answer after "####".
Bobby's dog needs to walk 2 minutes for every meal he eats.

$$2 \times 7 = 14 \text{ minutes.}$$

#### 14

## E Prompts

We listed the prompts we used in this work.

**Prompt for city recommendation**

Prompt = "Recommend a city to visit along with a reason. Respond ONLY with valid JSON with the following schema. Do not include explanations or extra text.
{
"city": "<city name>",
"reason": "<reason>"
}"

**Prompt for GSM8K**

Prompt = "{QUESTION} Please reason step by step, and put your final answer within \\boxed{}."

**Prompt for CNN-dailymail**

Prompt = f"Summarize the following article:
{ARTICLE}
Summary:"

**Prompt for Dolly**

prompt = f"### Instruction:
{INSTRUCTION}
### Input:
{INPUT}
### Response:"

**Prompt for Common-gen**

Prompt = f"Write a short story of 3–4 sentences that includes the following keywords: {KEY WORD}. The story should be coherent, creative, and naturally incorporate all the keywords.
Story:"

**Prompt of LLM as a diversity judge**

Instructions="The following are two files, each records responses generated by a model. The 'prompt' field stores prompts, and the 'response' field stores the corresponding groups of generated responses for the prompts. Please judge two models in terms of the diversity of their generated responses. You should give each model a score between [0,1]."
Input= f"Files: model 1: {FILE}; model 2: {FILE}."

# F LLM as a diversity judge

We use advanced LLMs (GPT-4o-mini) to score the diversity of the responses generated by the models. The following are the evaluations given by GPT-4o-mini. For fair comparison, we did two comparisons by changing the order of two models and taking the average scores.

**Diversity judged by gpt-4o-mini on GSM8K; Model 1 (GRPO), Model 2 (DQO)**

To evaluate the diversity of responses generated by the two models, we can analyze the patterns seen in the "responses" field of each model.
**Model 1:**

- **Characteristics:**
  - The majority of responses are quite repetitive, consistently using phrases like "To solve this problem, we'll break it down" or variations thereof.
  - There appears to be a lack of varied strategies or unique approaches in the generated answers, with many instances simply restating the problem-solving process in a similar structure.
- **Diversity Score:** Given the high level of similarity and repeated phrasing, I would rate the diversity of responses for Model 1 as **0.2/1**.

**Model 2:**

- **Characteristics:**
  - This model exhibits a somewhat greater variety in response structure, with different steps highlighted in the problem-solving process.
  - Phrases like "Step X: Determine the..." and "Let's denote the total number..." introduce variation, and there seems to be a more thoughtful approach to several responses.
  - While there are still some repeated forms, the presence of distinct steps and strategies indicates a broader array of response possibilities.
- **Diversity Score:** The variety and different methodologies present in the responses lead me to rate Model 2's diversity as **0.5/1**.

**Summary:**

- **Model 1 Score:** 0.2/1
- **Model 2 Score:** 0.5/1

Model 2 demonstrates better response diversity compared to Model 1, which relies heavily on similar language and structure.

Diversity judged by gpt-4o-mini on CNN-dailymail; Model 1 (PPO), Model 2 (DQO)

To assess the diversity of the responses generated by the two models, we evaluate the variety and uniqueness of the content provided in the responses to identical prompts.
**Model 1:**

- The responses from Model 1 vary in terms of length and detail, but many responses are quite structured and similar in approach.
- There is some variation in naming conventions and specific details, but overall, a noticeable pattern exists in how the information is expressed.
- Many responses use similar phrasing, leading to less diversity overall.

**Model 2:**

- Responses from Model 2 exhibit greater variation in phrasing, structure, and detail.
- The content shows a wide range of descriptions and perspectives, which enhances the uniqueness of each response.
- There is a noticeable effort to vary not just the wording but also the structure of each response, leading to a more diverse set of outputs.

**Scores:** Based on this analysis, we assign the following scores for diversity in generated responses:

- **Model 1: 0.4** (Limited diversity, responses similar in structure and phrasing)
- **Model 2: 0.8** (High diversity, significant variation in responses)

These scores reflect the extent of variety present in the responses generated by each model.

**Diversity judged by gpt-4o-mini on Dolly; Model 1 (PPO), Model 2 (DQO)**

To assess the diversity of responses generated by both models, we analyze the responses across various prompts, looking for variances in creativity, structure, and content diversity.

**Model 1 Analysis:**

- The responses from Model 1 often maintain a similar structure, primarily focusing on straightforward factual information, lists, or direct answers to prompts.
- The responses appear to have consistent formatting, providing succinct and clear answers, but can feel repetitive and lack creative exploration.
- While some responses do reflect a variety of information (e.g., multiple examples, details), the overall tone and presentation remain quite uniform.

**Model 2 Analysis:**

- Model 2 demonstrates a greater variety of response styles, incorporating markdown-like formatting, lists, and varying degrees of detail and structure across its outputs. This format enhances readability and introduces an engaging style.
- The responses in Model 2 often include additional editorial commentary or creativity (e.g., humorous tones in astrology interpretations, narrative elements in product recommendations).
- The use of diverse sentence structures and clear formatting contributes to a sense of dynamism in responses.

**Scoring:**

- **Model 1:** 0.4 (due to consistent structure and less creative diversity in responses)
- **Model 2:** 0.8 (due to varied response formats, creative styles, and more engaging outputs)

**Final Scores:**

- **Model 1:** 0.4
- **Model 2:** 0.8

**Diversity judged by gpt-4o-mini on Common-gen; Model 1 (DQO), Model 2 (PPO)**

To evaluate the diversity of the generated responses from both models, we analyze the content and variety of the responses provided for the same prompts.

**Model 1:**

- The responses feature varied character names and professions, indicating a range of different story premises.
- While the structure is somewhat consistent (retired individuals reflecting on their past), each response introduces a new character with unique backstories or contexts.
- The diversity is primarily in the character names and their respective professions, leading to different narrative angles.

**Model 2:**

- The responses are more repetitive in structure; most of them start similarly with "An older woman pauses...".
- There is a notable lack of variation in terms of settings and character actions across the responses. Although some details differ (such as the specific verb or object), the overall premise and sentence structure remain largely the same.
- This limits the diversity of storytelling in comparison to Model 1.

**Scores:**

- **Model 1:** 0.8 — Scores high for its unique character introductions and storytelling approaches, exhibiting good diversity.
- **Model 2:** 0.4 — Scores lower due to the repetitive structure and similarity in responses, which significantly reduces the diversity in storytelling.

**Summary:** Model 1 demonstrates a greater range of ideas and creativity in responses, while Model 2 lacks variety, leading to a more uniform storytelling style.