Enhancing Diversity in Large Language Models via Determinantal Point Processes

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Supervised fine-tuning and reinforcement learning, while improving large language model (LLM) quality, often reduce output diversity, leading to narrow, canonical responses. Existing methods to enhance diversity are limited, either by operating at inference time or by focusing on lexical differences. We propose a novel training method based on determinantal point processes (DPPs) to jointly optimize LLMs for quality and semantic diversity. Our approach samples and embeds responses, then uses the determinant of a kernel-based similarity matrix to measure diversity as the volume spanned by the embeddings. Experiments across instruction-following, story generation, and reasoning tasks demonstrate that our method substantially improves semantic diversity without sacrificing model quality.

1 Introduction

Post-training methods like supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) [32, 28, 22, 4] improve LLM quality but often sharply reduce output diversity [14, 20, 3, 29, 5]. Models trained this way tend to converge on narrow, canonical responses [14, 11], which is undesirable in settings like reasoning or personalization, where diverse outputs support multiple problem-solving approaches and user preferences.

Efforts to encourage diversity in LLM outputs mostly rely on inference-time methods like temperature scaling [1], top-k sampling [13], or related strategies [21, 10]. These approaches remain constrained by the base model's learned distribution. A stronger alternative is to optimize for diverse, high-quality outputs during training, but this poses two key challenges: (1) defining and optimizing diversity in a computationally efficient, theoretically grounded way, and (2) balancing diversity with response quality. Several recent works have attempted to enhance diversity in LLMs through training, but their efforts largely remain confined to the lexical level of diversity [31, 16, 17]. Encouraging the promotion of local lexical differences rather than enabling LLMs to produce a set of responses spanning distinct and meaningful modes of the answer space. Most related to our work, Chung et al. [6] propose a variant of DPO [22] by using pairwise distances in an embedding space. However, as we discuss in more detail in Section 3, pairwise distances provide a less robust and weaker notion of diversity.

In this work, we propose a principled training method based on determinantal point processes (DPPs) [15] to jointly optimize LLMs for quality and diversity. Unlike token-level entropy or lexical perturbations, our approach operates semantically: for each prompt, we sample responses, embed them with a pretrained encoder, and compute a kernel-based similarity matrix. The determinant of this matrix defines diversity as the volume spanned by the embeddings, while response rewards scale the vectors to balance quality and diversity. Experiments on instruction-following, story generation, and reasoning show that our method substantially improves semantic diversity without sacrificing quality.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Preliminaries

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 $I_k \in R^{k \times k}$ to represent the identity matrix with size k.

Determinantal point processes (DPPs). In this work, we quantify the diversity based on the concept of DPPs. We introduce the definiton of L-ensembles below which is a subclass of DPPs. For a comprehensive introduction to DPPs, please refer to Kulesza et al. [15].

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 semidefinite 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.*

The probability measure of DPPs inherently discourages the selection of similar items. If we think of the entries of L as measurements of similarity between pairs of elements, the determinant $\det(L_A)$ corresponds to the squared 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.

3 Proposed Methodology

Based on the above definition, given a set of responses $y_{1:k}$, we can formulate the diversity in this group of responses as $\mathrm{Div}(y_{1:k}) = \det(L_\phi(y_{1:k}))$ 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. In this work, we select the kernel function as the dot product, $f(\phi(y_i), \phi(y_j)) = \langle \phi(y_i), \phi(y_j) \rangle$. For simplicity, when it is clear from the context, we will omit the subscript in L_ϕ.

Our diversity definition has two advantages: it operates in embedding space, capturing semantic diversity, and its determinant-based formulation measures group rather than pairwise diversity. Pairwise metrics, like average distances, are prone to the "clustering" phenomenon, where a few separated groups create a false sense of diversity [25]. In contrast, the determinant rewards linearly independent responses, penalizing clusters and low-dimensional subspaces. This encourages exploration of the full embedding space, ensuring genuine semantic diversity.

Reinforcement learning has been a popular method for post-training LLMs with either an existing reward function or the one inferred from a preference dataset, i.e. RLHF. With the reward function, the model is optimized by maximizing the following KL-regularized objective,

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

where $J(\pi_\theta) = \mathbb{E}_{x, y \sim \pi(\cdot|x)}[r(x, y)]$ is the expected return and β is a hyperparameter balancing the KL divergence and rewards. As we have pointed out in the introduction section, after alignment, the model tends to converge toward a narrow set of responses, leading to limited diversity. To solve this issue of diversity collapse, inspired by the concepts from DPPs, we propose the following objective to directly optimize LLMs for both quality and diversity in generated responses,

$$J_{Div}(\pi_\theta) = \mathbb{E}_{x, y_1, \ldots, y_k \sim \pi_\theta(\cdot|x)} \left[\sum_{i=1}^{k} r(x, y_i) + \alpha \log \det(L_\phi(y_{1:k})) \right]. \tag{2}$$

For each prompt, we sample k responses $y_{1:k}$ from the model like what people do with Group Relative Policy Optimization (GRPO). Instead of just optimizing the reward, we add a diversity term as part of the objective. And α is a hyperparameter to balance the quality and diversity.

It can be shown that by optimizing (1) with our $J_{Div}(\pi_\theta)$, the optimal policy satisfies,

$$\pi_{div}(y_{1:k}|x) \propto \pi_{ref}(y_{1:k}|x) \exp\left(\frac{1}{\beta} \left(\sum_{i=1}^{k} r(x, y_i) + \alpha \log \det(L_\phi(y_{1:k})) \right) \right) \tag{3}$$

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

$$\pi_{div}(y_{1:k}|x) \propto \det(L_\psi(y_{1:k})) \tag{4}$$

The above expression tells us that our policy (4) learns to generate a group of responses with the probability proportional to the determinant of the gram matrix formed by the embedding vectors of these responses. From a geometric view, our policy can pick a group of vectors in the embedding space of responses according to the squared volume of the space spanned by these vectors.

3.1 Algorithm

We noticed that implementing the objective (2) in practice poses several challenges, such as the high variance of the gradient estimator and potential numerical explosion. We present a practical version of the algorithm designed to stabilize training. The gradient of $J_{Div}(\pi_\theta)$ can be calculated as follows,

$$\nabla J_{Div}(\pi_\theta) = \mathbb{E}_{x, y_{1:k} \sim \pi_\theta(\cdot|x)} \left[\sum_{i=1}^{k} \nabla \log \pi_\theta(y_i|x)(r(x, y_i) + \alpha \log \det(L(y_{1:k}))) \right] \tag{5}$$

The first issue is the determinant of $L(y_{1:k})$ can be close to zero which leads to a super negative value of $\log(\det(L(y_{1:k})))$. The unboundedness of the diversity term makes the training process unstable and also makes the balance between quality and diversity difficult such that only a carefully selected α is effective. To fix this issue, we propose to consider the determinant of the matrix $L(y_{1:k}) + I_k$. By adding an identity matrix, we can show $k \geq \log(\det(L(y_{1:k}) + I_k)) \geq 0$.

The second issue is that the gradients is the summation of the gradient of k responses. It has high variance especially when k is large. 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,

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

The *loo* estimator is unbiased and has the following nice property.

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* $\log(1 + k) \geq \log \frac{\det(L(y_{1:k})+I)}{\det(L(y_{-i})+I)} \geq 0$.

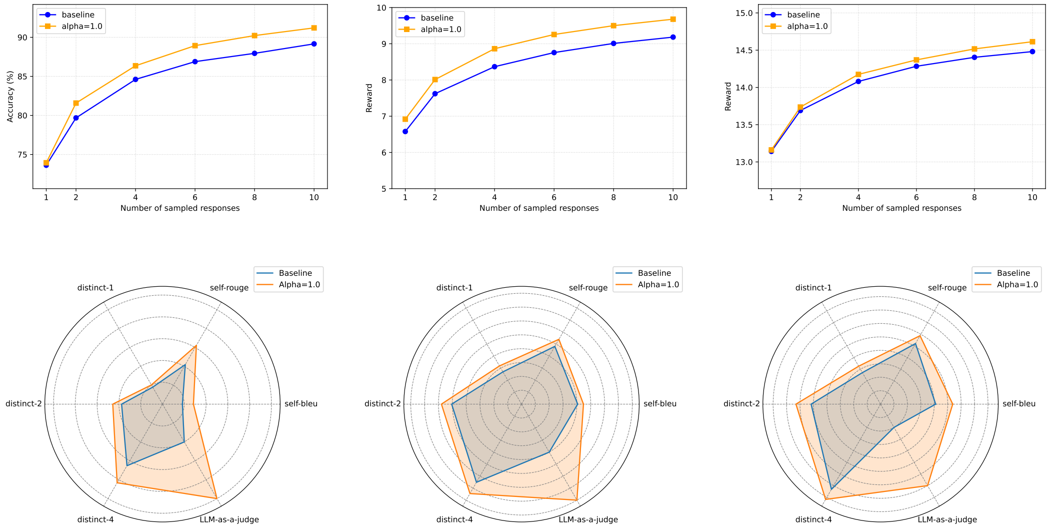
4 Experiments

We run experiments under three different kinds of tasks, including reasoning (GSM8K [7]), story-writing (Common-Gen [18]) and instruction-following (Dolly [8]). We compare our algorithm to the baseline which trains the model with only reward. For the detailed experimental setup, please see Appendix C.

We use $pass@n$ metric to measure the quality with n varies from 1 to 10. 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 and Self-ROUGE 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 D and E).

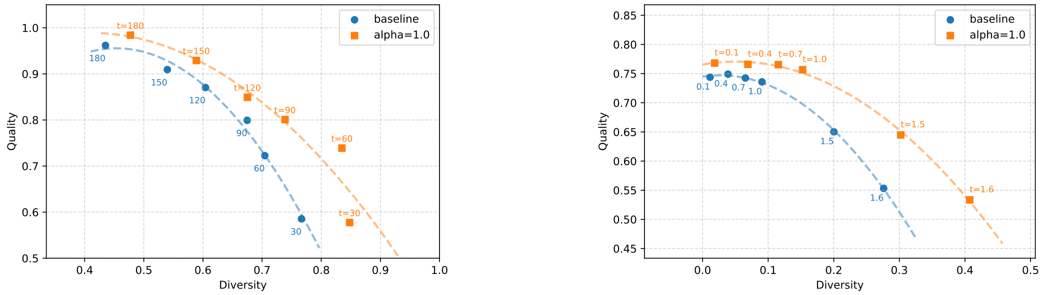
Quality In Figure 1, we show the $pass@n$ performance across three tasks. We compare the baseline model trained with only reward and the model trained by our objective with hyperparameter $\alpha = 1$. Our model exhibits better performance than the baseline model especially when n is large. Besides, 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) Dolly (c) Gen

Figure 1: The performance on $pass@n$ and diversity metrics. Baseline: the model trained with only reward; Alpha=1.0: ours.

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 1. For each metric, higher values indicate greater diversity. As shown in the figure, the model trained with our method consistently outperforms the model trained solely with reward, 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 E), highlighting improvements at the semantic level.



(a) Gen (b) GSM8K

Figure 2: 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.

Pareto frontier To illustrate how our model achieves a favorable balance between quality and diversity, we plot the Pareto frontiers of our model and the baseline model by varying either the training steps or the sampling temperature in Figure 2. Across different sampling temperatures (the right in Figure 2), 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 2), 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.

References

[1] 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`.

[2] 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.

[3] 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.

[4] 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`.

[5] 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.

[6] 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`.

[7] 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.

[8] 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`.

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

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

[11] Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[12] 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`.

[13] 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`.

[14] 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`.

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

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

[17] 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.

[18] 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`.

[19] 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.

[20] Sonia K. Murthy, Tomer D. Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *NAACL (Long Papers)*, pages 11241–11258, 2025. URL `https://doi.org/10.18653/v1/2025.naacl-long.561`.

[21] 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.

[22] 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`.

[23] 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`.

[24] 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.

[25] 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.

[26] 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`.

[27] 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`.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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 $\psi(x, y) = \sqrt{\exp(\frac{r(x,y)}{\beta})\pi_{ref}(y|x)} \cdot \phi(y)$, then the optimal policy in (3) 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

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

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 $\alpha = \beta$,

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

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 selected feature vectors, which enforces strict linear independence: any subset that induces a singular L 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 σ^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 [9]. 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 $\lambda_1 \leq \cdots \leq \lambda_n$ and B has eigenvalues $\beta_1 \leq \cdots \leq \beta_m$. Then,

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

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

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

Proof. We use the Courant–Fischer min–max theorem. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, 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 $\beta_1 \leq \cdots \leq \beta_m$, 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

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

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

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

which is exactly the interlacing property.

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

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

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

and,

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

Since $L(y_{1:k})$ is positive semidefinite, it holds $\lambda_i \geq 0$, $\forall i$. 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 $k \geq \lambda_k \geq \lambda_1 \geq 0$.

B Related works

Evaluating Diversity of LLMs. Several works have focused on evaluating the diversity of LLM generated content [12, 26], also on investigating the impact of post-training on diversity metrics [14, 27]. The lack of diversity in LLM generated content also affects text written by humans using LLMs [23].

Improving Diversity of LLMs. There are mainly two lines of works on promoting diversity in LLMs. One focuses on inference strategies. Nguyen et al. [21] 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 [10], considers the largest difference between consecutive probabilities of tokens in a sorted distribution to promote diversity while maintaining correctness. Ahmed et al. [2] 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 work focuses on the training strategy to best elicit diversity from LLMs. Lanchantin et al. [16] 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. [31] 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. [17] 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. [6] 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) [15], 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. [24] proposed a DPPs-based algorithm to train a population of diverse polices in reinforcement learning for better exploration.

C Experimental setup

Data preparation For GSM8K dataset, we directly use the training and test split. 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 [30] 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 [19] and embedding model sentence-transformers/all-MiniLM-L6-v2.

Table 1: Training configurations. For max prompt and response length, we use different values for different datasets. From left to right, it corresponds to GSM8K, 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, 1024, 64
max response length	256, 512, 128

D Prompts

We listed the prompts we used in this work.

Prompt for GSM8K

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

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}."

E 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.

Diversity judged by gpt-4o-mini on GSM8K; Model 1 (baseline), Model 2 (ours)

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 Dolly; Model 1 (baseline), Model 2 (ours)

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 (ours), Model 2 (baseline)

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.