DIVE: Diversified Iterative Self-Improvement

Yiwei Qin[2] **and Yixiu Liu**[1,2] **and Pengfei Liu**[1,2*]
[1]Shanghai Jiao Tong University
[2]Generative AI Research Lab (GAIR)
{qinyiwei07@outlook.com, pengfei@sjtu.edu.cn}

Abstract

Recent advances in large language models (LLMs) have demonstrated the effectiveness of Iterative Self-Improvement (ISI) techniques. However, continuous training on self-generated data leads to reduced output diversity, a limitation particularly critical in reasoning tasks where diverse solution paths are essential. We present DIVE (Diversified Iterative Self-Improvement), a novel framework that addresses this challenge through two key components: Sample Pool Expansion for broader solution exploration, and Data Selection for balancing diversity and quality in preference pairs. Experiments on MATH and GSM8k datasets show that DIVE achieves a 10% to 45% relative increase in output diversity metrics while maintaining performance quality compared to vanilla ISI. Our ablation studies confirm both components' significance in achieving these improvements. Code is available at https://github.com/qinyiwei/DIVE.

1 Introduction

Recent advancements in large language models (LLMs) have driven significant improvements through self-improvement techniques (Zelikman et al., 2022; Madaan et al., 2024; Wang et al., 2022), where models enhance their capabilities by refining their performance based on feedback, often using their own outputs for further enhancement. Two prominent approaches in this area are Reinforcement Learning (RL) (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022) and Preference Learning (Rafailov et al., 2024; Zhao et al., 2023; Ethayarajh et al., 2024; Azar et al., 2024), both of which enable models to refine their behavior by optimizing for feedback signals, such as rewards or preferences. Iterative Self-Improvement (ISI) extends these methods by using an iterative process, where models continuously leverage previous outputs to generate more refined responses, proving highly effective in various domains from general instruction-following (Xu et al., 2023; Yuan et al., 2024) to specialized areas like mathematical reasoning (Pang et al., 2024; Mitra et al., 2024).

Despite the positive outcomes of ISI in enhancing model performance, recent research has identified model collapse as a critical challenge when training models on self-generated data (Shumailov et al., 2024; Dohmatob et al., 2024; Gerstgrasser et al., 2024). This phenomenon, where models progressively lose information about the underlying distribution, is particularly relevant to ISI processes as models continuously learn from their own outputs. In RL and preference learning settings, this issue manifests as reduced diversity in generated responses, as the model increasingly focuses on a narrow set of high-reward patterns (Wu et al., 2024; Kirk et al., 2023).

While recent advancements in reasoning with LLMs have focused on improving accuracy through top-ranking solutions, they often overlook the importance of diverse reasoning paths. Methods like Self Consistency (Wang et al., 2022), ToT (Yao et al., 2024) and RAP (Hao et al., 2023) rely on the LLM's capacity to explore diverse reasoning solutions, leveraging the intuition that complex reasoning tasks typically admit multiple valid paths to the correct answer (Evans, 2010; Stanovich, 2012). Although some studies have investigated techniques to enhance reasoning diversity (Wang et al., 2022; Xie et al., 2024; Li et al., 2022; Naik et al., 2023; Yu et al., 2024), the challenge of diversity loss in ISI remains underexplored.

To address this challenge, we present **D**iversified **I**terative Self-Impro**VE**ment (DIVE), shown in Fig.1, the first study focused on this problem. DIVE operates through two complementary strategies in the preference learning stage: (1) Sample Pool Expansion and (2) Data Selection. Sample Pool Expansion encourages the model to explore a broader set of potential solutions at each iteration

[*] Corresponding author

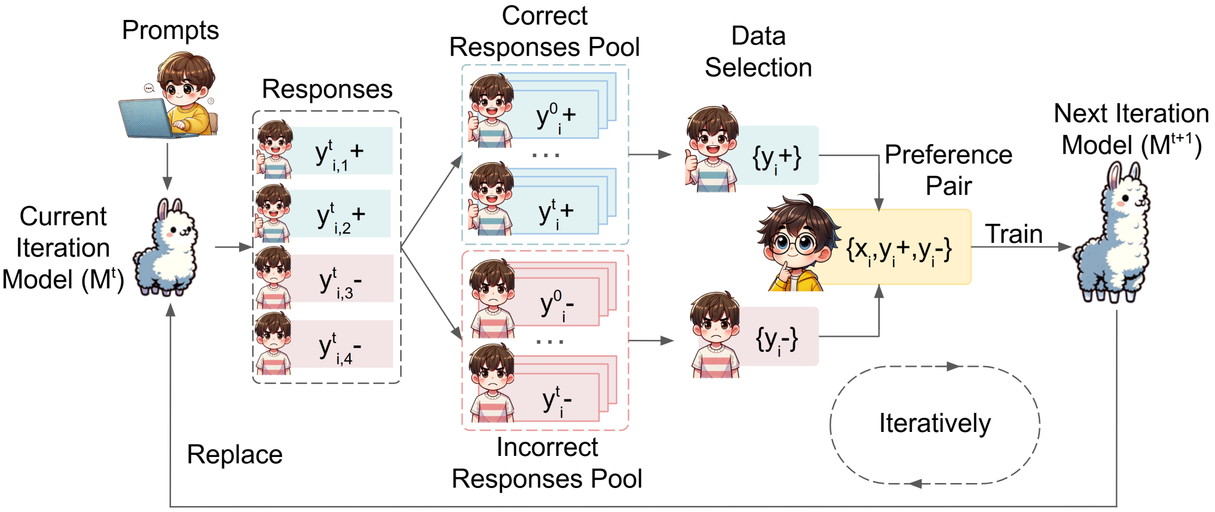

Figure 1: Overview of the Diversified Iterative Self-Improvement (DIVE) framework. At each iteration t, the process includes response generation, pool expansion through correct and incorrect response collection, data selection for balancing quality and diversity, and model refinement through preference learning, producing an improved model M^{t+1} for the next iteration.

by sampling more responses per question and incorporating data from all previous iterations. Data Selection then applies outlier detection techniques to filter responses for quality while using greedy selection algorithms to maximize diversity in the preference pairs. By curating diverse yet high-quality preference pairs, DIVE guides the model to generate varied outputs while maintaining performance.

Our experimental results demonstrate that DIVE significantly enhances the diversity of model outputs on the MATH and GSM8k datasets compared to vanilla ISI, achieving a 10% to 45% relative increase across various diversity metrics for both positive and negative examples, without compromising output quality. Ablation studies further highlight the critical roles of Sample Pool Expansion and Data Selection in driving these results.

2 Methodology

Let $D = (x_i, y_i)_{i=1}^N$ represent a training set containing questions x_i and their corresponding ground truth response y_i. We begin with a foundation model, typically a pre-trained model denoted as M_{PT}. The objective of self-improvement is to enhance the M_{PT}'s performance by refining its capabilities using its own outputs, without relying on external signals. When this process is repeated over multiple training rounds, it becomes ISI, where the model incrementally improves by applying preference learning to its own generated responses at each iteration.

2.1 Iterative Self Improvement

Direct Preference Optimization (DPO) (Rafailov et al., 2024) DPO is a widely-used method for offline preference learning that enables direct optimization of model preferences without requiring an explicit reward model. The key insight of DPO is to express the probability of preference data using the ratio between the policy model and a reference model. The DPO objective is defined as:

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y^+,y^-)\sim D_{\text{pref}}} \left[\log \sigma(r)\right],$$
$$r = \beta \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \beta \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)} \quad (1)$$

where (x, y^+, y^-) represents preference pairs from the preference dataset D_{pref}, with x being the input question, y^+ the preferred (correct) response, and y^- the non-preferred (incorrect) response. The policy model π_θ learns to assign higher probability to preferred responses compared to non-preferred ones.

To stabilize the DPO training and prevent the model from deviating too far from its initial behavior, we incorporate an additional negative log-likelihood (NLL) loss on the chosen sequences (Pang et al., 2024; Dubey et al., 2024; Xu et al., 2024). This helps maintain response consistency while allowing for targeted improvements through preference learning. The NLL loss term is defined as:

$$L_{\text{NLL}} = -\mathbb{E}_{(x,y^+)\sim D_{\text{pref}}} \frac{\log \pi_\theta(y^+|x)}{|y^+|} \quad (2)$$

The final loss function combines the DPO and NLL losses as follows:

$$L_{\text{pref}} = \alpha \cdot L_{\text{DPO}} + (1-\alpha) \cdot L_{\text{NLL}} \quad (3)$$

where α is a hyperparameter that balances the contributions of DPO and NLL losses.

Iterative Training We start by performing supervised fine-tuning (SFT) on the pre-trained model M_{PT} using dataset D, producing a fine-tuned model M_0. In ISI, a series of models $M_1, \ldots, M_T$ are trained, where each model M_t builds upon the outputs of the previous model M_{t-1}. During each iteration, preference data for training M_t is sampled from M_{t-1}, and M_{t-1} is used as the reference model in the DPO loss. The steps for each iteration are as follows:

1. **Data Sampling**: In the t-th iteration, for each question $x \in D$, we sample K responses from the model M_{t-1} to form the candidate pool: $D^t_{\text{pool}} = \{(x_i, y^j_i)|x_i \in D, j \in [1, K]\}$.

2. **Preference Pair Construction**: The candidate pool D^t_{pool} is divided into a correct pool D^{t+}_{pool} and an incorrect pool D^{t-}_{pool} by comparing the generated response with the gold-standard answer. If the final answer of a generated response matches the gold standard, the response goes to D^{t+}_{pool}; otherwise, it goes to D^{t-}_{pool}. From these pools, we select P responses to construct the preference dataset: $D^t_{\text{pref}} = \{(x_i, y^+_i, y^-_i)|x_i \in D, y^+_i \in D^{t+}_{\text{pool}}, y^-_i \in D^{t-}_{\text{pool}}\}$.

3. **Preference Training**: Using the preference dataset D^t_{pref}, the model M_{t-1} is refined into M_t by optimizing the preference loss L_{pref}.

2.2 Diversified Iterative Self-Improvement

As highlighted in Wu et al. (2024); Kirk et al. (2023), preference learning often leads to a reduction in diversity, a problem that is exacerbated in iterative settings due to the accumulation of this effect over time. We propose two complementary strategies to address this challenge: Sample Pool Expansion, which enlarges the candidate pool for response selection, and Data Selection, which ensures diverse yet high-quality examples are chosen for training. These strategies work within the existing ISI framework while effectively maintaining output diversity.

2.2.1 Sample Pool Expansion

To provide more candidates for constructing diverse preference pairs, we expand the candidate sample pool D_{pool} through two complementary strategies. A larger sample pool offers more options for the subsequent data selection process, which is crucial for selecting diverse examples for preference learning.

Increased Sampling per Question At each iteration, we increase the number of responses K sampled per question, providing a broader set of candidates for preference learning.

Global Data Usage Instead of relying solely on the responses generated by model M_{t-1} for training M_t, we incorporate global data from all previous iterations. This expanded pool is defined as $D^t_{\text{pool}} = \bigcup_{i=1}^{t} D^i_{\text{pool}}$ ensuring that no information from previous iterations is lost and avoiding extra sampling computation.

2.2.2 Data Selection

Our preliminary experiments show that the diversity of the examples selected for preference learning, rather than the overall diversity of the response pool, significantly impacts the model's ability to generate diverse outputs after training. Thus, it is crucial to carefully select diverse examples from the response pool for preference learning.

Greedy Selection Method We use a greedy algorithm to maximize the diversity of the selected responses, following these steps:

1. Randomly select one response from D_{pool} and add it to the selected response list. Remove this response from D_{pool}.

2. For each remaining response in D_{pool}, calculate the diversity of the selected response list as if the current example were added.

3. Select the response that maximizes the diversity of the selected list, add it to the list, and remove it from D_{pool}.

4. Repeat Steps 2 and 3 until either $D_{\text{pool}} = \varnothing$ or the desired number of responses P is reached.

While this method increases diversity effectively, we observed that focusing solely on diversity can negatively impact model accuracy. We hypothesize that maximizing diversity may lead to selecting low-quality, outlier responses that harm the model's performance.

Balancing Quality and Diversity To mitigate this issue, we first filter the response pool using the Isolation Forest method (Liu et al., 2008), with features derived from Sentence-BERT embeddings (Reimers, 2019) that capture the semantic aspects of the responses. Using distances in the embedding space, we identify and exclude extreme outliers (responses that deviate significantly from the general distribution of valid solutions) to maintain response quality.

Once the response pool is filtered, we apply the greedy selection method to maximize diversity among the remaining high-quality responses. This ensures a balanced selection process that maintains both diversity and quality in the final model.

3 Experiment

3.1 Experimental Settings

3.1.1 datasets

We conducted experiments on two math reasoning datasets:

GSM8K (Cobbe et al., 2021): This dataset contains grade-school-level math word problems. Each problem consists of a question x_i and a solution y_i, which includes a gold chain-of-thought (COT) explanation (Wei et al., 2022) and a final numerical answer. The training set consists of 7,473 examples, and the test set contains 1,319 examples.

MATH (Hendrycks et al., 2021): This dataset contains more advanced math problems. Similar to GSM8K, each problem provides a gold CoT solution along with a final answer. The training set includes 7,500 problems, while the test set contains 5,000 examples.

In the self-improvement paradigm, for both datasets, we utilize only the questions from the training set for preference learning, without introducing any additional questions. The correctness of the model-generated solutions is judged based on the final answers provided in the gold solutions.

3.1.2 Evaluation Metrics

To assess how well the model balances quality and diversity, we adopt two types of evaluation metrics that measure performance from both aspects:

Quality For quality evaluation, we use the following metrics: **@1 Accuracy** which measures the model's accuracy when sampling a single response. It tests how well the model ranks the sample space, with a focus on whether the correct response is placed at the top-1 position. **@50 Accuracy** which evaluates the model's accuracy when sampling 50 responses. The model is considered correct if any of the 50 responses is correct. This metric tests the model's potential to solve a question when sampling more responses.

Diversity To evaluate the diversity of the generated responses, we use the following metrics, in line with Kirk et al. (2023): **Distinct N-grams** (Tevet and Berant, 2020) which counts the number of distinct N-grams (averaged over $n = 1, \ldots, 5$) in the set of outputs, which provides a measure of lexical diversity. **Sentence-BERT Embedding Cosine Similarity** (Li et al., 2015) which embeds each response using a Sentence-BERT model and calculates the average cosine similarity between the embeddings. The diversity score is then calculated as $1 -$ average similarity, where lower similarity indicates higher diversity. Both of these methods have been shown to align well with human evaluations of diversity (Tevet and Berant, 2020), enabling us to quantify the diversity of the model's outputs effectively.

3.1.3 Training Details

Our experiments are based on the pre-trained language model Mistral-7B (Jiang et al., 2023). For SFT, we fine-tune Mistral-7B on the GSM8K/MATH Train subset to produce the initial model, M_0. The fine-tuning is done using full-model fine-tuning with a learning rate of 1×10^{-6}, a cosine learning rate schedule, 3 epochs.

For the ISI phase, at each iteration t, we generate $K = 10$ or 50 solutions per question from the GSM8K/MATH Train subset to form the response pool D^t_{pool}, using nucleus sampling with top_$p = 0.95$ and temperature $T = 0.7$, based on the model M_{t-1}. For experiments without global data usage, $P = 5$ preference pairs are constructed from D^t_{pool}. For experiments with global data usage, the pool is expanded to $D^t_{\text{pool}} = \cup^t_{i=1} D^i_{\text{pool}}$. [1]

[1] Since some questions may have fewer than $P = 5$ correct or incorrect responses, we construct at most P preference pairs per question. Questions with no correct or no incorrect responses in the pool are skipped without constructing any

	Method	Dis-N Pos	Dis-N Neg	SentBERT Pos	SentBERT Neg	@1	@50
Sample 10	Vanilla	0.345	0.454	0.111	0.168	0.704	0.976
	Global	0.350	0.444	0.119	0.182	0.707	**0.980**
	Selection	0.388	0.462	0.125	0.184	0.703	0.975
	Global+Selection	0.397	0.507	0.132	0.196	0.707	0.975
Sample 50	Vanilla	0.309	0.380	0.106	0.168	0.718	0.975
	Global	0.348	0.462	0.118	0.184	0.716	0.974
	Selection	0.440	**0.538**	0.145	0.214	0.716	0.976
	Global+Selection	**0.448**	0.502	**0.152**	**0.224**	**0.722**	0.972

Table 1: Comparison of different diversity enhancement methods on GSM8k dataset using Mistral-7B as the base model. Results show diversity metrics (Dis-N and SentBERT) for both positive and negative examples, along with accuracy metrics. All metrics have been normalized so that higher values consistently indicate better performance. **Bold** indicates the best overall performance across all settings, while underline represents the best performance within their respective sampling group (Sample 10 or Sample 50).

	Method	Dis-N Pos	Dis-N Neg	SentBERT Pos	SentBERT Neg	@1	@50
Sample 10	Vanilla	0.647	0.557	0.247	0.304	0.176	0.580
	Global	0.636	0.550	0.242	0.300	**0.194**	**0.610**
	Selection	0.662	0.565	0.245	0.311	0.178	0.600
	Global+Selection	0.665	0.573	0.254	0.310	0.188	**0.610**
Sample 50	Vanilla	0.612	0.540	0.228	0.283	0.186	0.606
	Global	0.635	0.542	0.247	0.299	0.190	0.606
	Selection	**0.694**	**0.612**	0.264	0.313	0.188	0.594
	Global+Selection	0.692	0.599	**0.273**	**0.326**	**0.194**	0.586

Table 2: Results on MATH dataset with identical experimental settings as Table 1.

We run up to $T = 6$ iterations, producing models $M_1, M_2, \ldots, M_6$. In each iteration, we train for one epoch on all the preference pairs constructed so far, with the number of pairs per iteration ranging from 10k to 30k, depending on the setting. [2]

The loss coefficient α is set to 0.5, and the DPO coefficient β is set to 0.4. Full-model fine-tuning is used, with a batch size of 8, gradient accumulation steps of 2, and a learning rate of 3×10^{-8} using the AdamW optimizer with a constant learning rate schedule. Training is conducted on four A100 GPUs (80G memory) with a total batch size of 64.

3.2 Experimental Results

To evaluate the effectiveness of our proposed methods, we conduct experiments with two sampling sizes (10 and 50) comparing four variants of ISI:

1. Vanilla: The standard ISI method as our baseline
2. Global: Expanding sample pool with global data (Section 2.2.1)
3. Selection: Applying data selection for quality and diversity (Section 2.2.2)
4. Global + Selection: Combining both global data expansion and data selection

Tables 1 and 2 present the main results from the best-performing iteration (out of six) for each method. Our analysis reveals several key findings:

Quality Preservation. All three proposed methods (Global, Selection, and Global+Selection) maintain performance comparable to the baseline in terms of @1 and @50 accuracy on both GSM8K and Math datasets, demonstrating that our diversity-enhancing techniques do not compromise model quality.

Impact of Sampling Pool Size. With larger sampling size (50 vs 10), the vanilla method shows lower diversity, indicating that naive sampling expansion can actually harm diversity. Interestingly, the Global method alone does not consistently improve diversity over the vanilla baseline, suggesting that sample pool expansion without proper diversity management is insufficient.

Effectiveness of Data Selection. The data selection mechanism consistently enhances diversity across all settings (Sample 10/50, GSM8K/Math).

preference pairs.

[2] As model performance improves over iterations, fewer incorrect examples and more correct examples are generated, leading to varied number of preference pairs being constructed in each iteration.

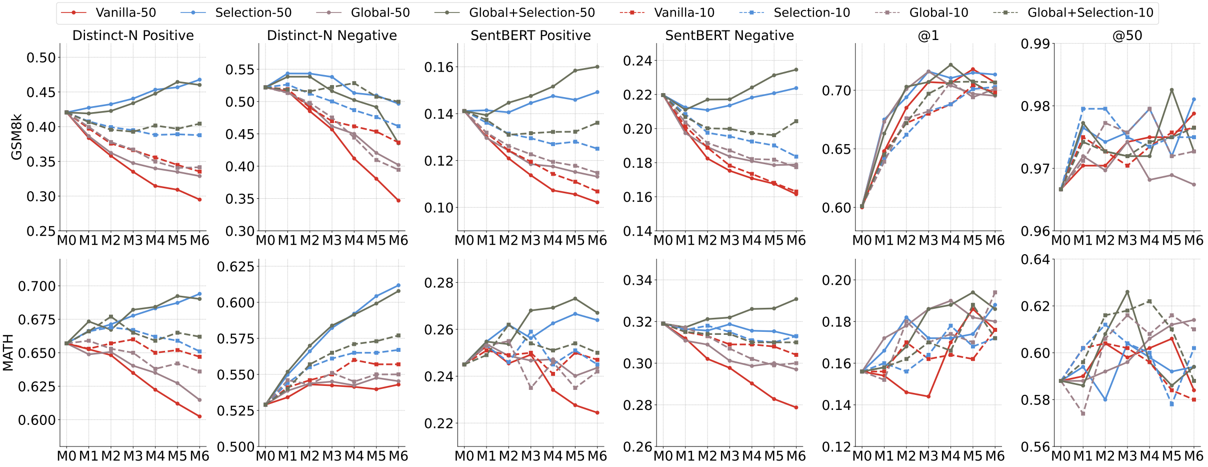

Figure 2: Evolution of diversity metrics and model performance across iterations (M0-M6) for both GSM8k and MATH datasets. Each subplot shows different evaluation metrics: Distinct-N for positive and negative examples, SentBERT embeddings similarity, and accuracy measures. Solid and dashed lines with different colors represent different sampling settings and methods.

This is evidenced by clear improvements from Vanilla to Selection and from Global to Global + Selection. Notably, the combination of large sampling (50) with Global + Selection achieves the highest diversity across most metrics.

Iterative Analysis. Figure 2 illustrates the dynamics across all six iterations:

1. **Diversity Evolution:** In vanilla ISI, diversity consistently declines across iterations, with larger sampling sizes (50) showing more severe reduction compared to smaller ones (10). Our Global + Selection method, in contrast, maintains and even improves diversity throughout iterations.

2. **Performance Trends:** All methods show accuracy improvements of 10-12 points on GSM8K and 2-4 points on Math, typically peaking between iterations 4-6 before saturation. The stable @50 accuracy across iterations suggests that self-improvement primarily acts as a re-ranking mechanism, consistent with observations in Wu et al. (2024).

3. **Sample Size Effects:** Larger sampling (50) yields marginally better accuracy and significantly higher diversity compared to smaller sampling (10), indicating that increased sampling, when properly managed, benefits both quality and diversity.

Ablation Analysis. Our experiments serve as an ablation study to validate each component's contribution. For data selection, the consistent superiority of Selection over Vanilla in diversity metrics demonstrates its effectiveness. For sample pool expansion, the advantage of Global + Selection over Selection, larger sampling (50) over smaller sampling (10), confirms the benefit of incorporating global data. These results verify that both components are essential for maximizing diversity while maintaining performance.

4 Analysis

To gain deeper insights into diversity challenges in ISI and evaluate the effectiveness of DIVE, we investigate three key questions: Q1: Can increasing the number of samples per question alone adequately substitute for using a global data pool to expand the sample set? Q2: How does question difficulty affect diversity throughout the iterative process? Q3: How robust are our diversity improvements across different evaluation metrics?

4.1 Impact of Global Data Usage (Q1)

While both global data accumulation and increased per-question sampling can expand the sampling pool size, their effectiveness may differ. To investigate this, we compare three approaches across six iterations: 1.Selection: the sampling pool size remains constant at 10-10-10-10-10-10. 2.Global+Selection: the sampling pool size expands incrementally to 10-20-30-40-50-60 when global data is included, as each iteration incorporates all previous ones. 3.Selection+Increased Sampling: the sampling pool size is 10-20-30-40-50-60 via increased sampling count.

As shown in Figure 3, while Selection+Increased Sampling shows improved diversity in later iterations, Global+Selection consistently achieves higher diversity across all metrics for both positive

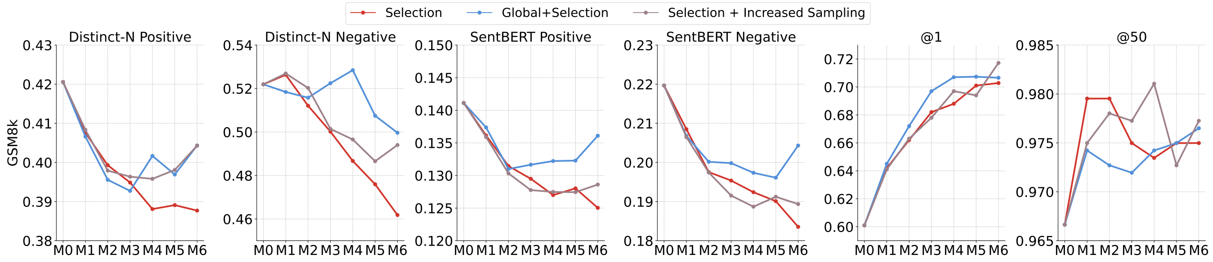

Figure 3: Comparison of different sampling strategies for GSM8k dataset.

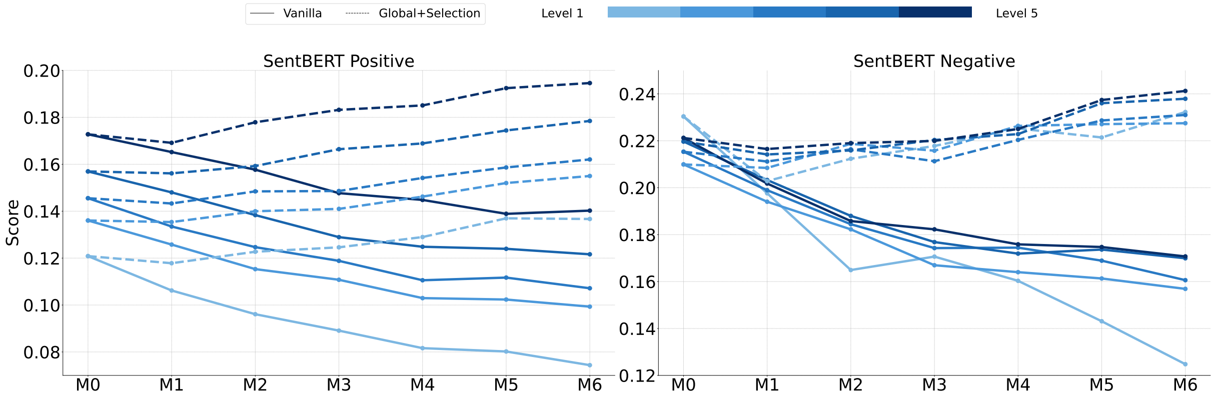

Figure 4: Diversity trends across different difficulty levels (Level 1-5) for positive and negative examples. The plots demonstrate how question difficulty influences output diversity during the ISI process.

and negative examples. This suggests that diversity lost in early iterations is difficult to recover through increased sampling alone, underscoring the importance of leveraging accumulated data. Moreover, Global+Selection achieves this with lower computational cost, requiring only 60 total samples per question compared to 210 for Selection+Increased Sampling, demonstrating both the effectiveness and efficiency of global data incorporation.

4.2 Diversity Across Difficulty Levels (Q2)

Our experiments on GSM8K and MATH datasets reveal an intriguing pattern: the more challenging MATH dataset maintains higher diversity and shows less pronounced diversity loss during ISI. This observation motivates us to investigate the relationship between question difficulty and diversity patterns. To systematically analyze this relationship, we classify questions into five difficulty levels based on their correct ratio R (percentage of correct answers when sampling 50 examples)[3]. This automated approach enables objective difficulty assessment without manual annotation.

As shown in Figure 4, our analysis reveals several key findings: 1.Difficulty-Diversity Correlation: Harder questions consistently exhibit higher diversity in positive examples, though this correlation is less pronounced for negative examples. 2.Differential Diversity Loss: Easier questions suffer more severe diversity loss during iteration (e.g., Level 1 shows 53.4% and 43.8% drops for negative and positive examples respectively, compared to 25.0% and 19.7% for Level 5) 3.Method Robustness: DIVE demonstrates consistent diversity improvements across all difficulty levels, indicating its effectiveness is not biased toward any particular difficulty range.

4.3 Alternative Metrics for Diversity (Q3)

To validate the robustness of our diversity improvements, we extend our evaluation beyond the metrics in Section 3.1.2, incorporating both advanced embedding-based and task-specific metrics.

Advanced Embedding Metrics We employ two state-of-the-art embedding models for diversity assessment: **NV-Embed** (Moreira et al., 2024)[4]: A 7B parameter model currently leading the MTEB

[3] Difficulty levels are defined as Level 5 (hardest): $0 \leq R < 0.2$; Level 4: $0.2 \leq R < 0.4$; Level 3: $0.4 \leq R < 0.6$; Level 2: $0.6 \leq R < 0.8$; Level 1 (easiest): $0.8 \leq R \leq 1$

[4] Available at https://huggingface.co/nvidia/NV-Embed-v2

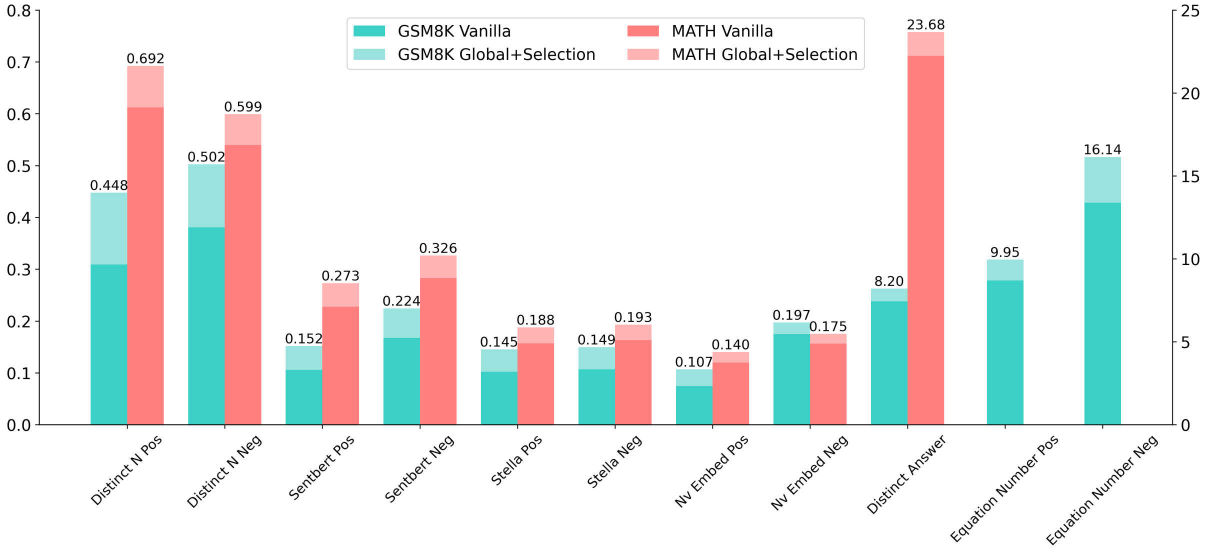

Figure 5: Results of different diversity metrics for both the GSM8k and MATH datasets. Only the results from the iteration with the highest accuracy are shown, while the results for all iterations are provided in Appendix A.1.

Leaderboard (Muennighoff et al., 2022). **Stella** [5]: The top-performing 1.5B parameter model on MTEB.

Mathematical Reasoning Metrics We introduce two metrics specifically designed to capture diversity in mathematical reasoning: **Distinct Equation Chains**: This metric counts the number of distinct equation sequences in model-generated solutions, where each sequence represents a unique reasoning path.[6] **Distinct Answers**: Counts unique final answers per question, primarily reflecting diversity in incorrect solutions as correct answers are consistent.

As shown in Figure 5, Global+Selection demonstrates consistent improvements across all eleven diversity metrics. Notably, while our method uses computationally efficient metrics (SentBERT and Distinct-N) during training, the improvements generalize to more sophisticated metrics, confirming the robustness of our approach.

5 Related Work

Diversity in Reasoning Research on diversity in language models has evolved from general text generation diversity (Batra et al., 2012; Li et al., 2016; Vijayakumar et al., 2018) to the specific challenges of reasoning tasks, where the goal is to generate diverse yet valid solution paths. Recent work has explored various approaches: Wang et al. (2022) demonstrate that sampling multiple reasoning paths improves answer accuracy through aggregation, while Xie et al. (2024) combines beam search with temperature sampling to balance quality and diversity. Other approaches include varying prompts to enhance solution diversity (Li et al., 2022), using model feedback to encourage multiple solving strategies (Naik et al., 2023), and modeling reasoning as a Markovian flow for diverse path generation (Yu et al., 2024).

Iterative Self-Improvement Recent advances in ISI have shown promising results in enhancing model capabilities through self-play and iterative refinement, particularly in mathematical reasoning (Pang et al., 2024; Mitra et al., 2024; Wu et al., 2024). However, when models are trained on self-generated data, they may experience model collapse, where models progressively lose information about the underlying distribution (Shumailov et al., 2024; Dohmatob et al., 2024; Gerstgrasser et al., 2024). This phenomenon has been observed in various settings including preference learning methods like DPO (Rafailov et al., 2024) and RLHF (Ouyang et al., 2022), where it manifests as reduced output diversity (Kirk et al., 2023; Wu et al., 2024). While existing work suggests maintaining a balanced mix of human-authored and model-generated data to preserve model performance (Shumailov et al., 2024; Dohmatob et al., 2024; Gerstgrasser et al., 2024), our work intro-

[5] Available at `https://huggingface.co/dunzhang/stella_en_1.5B_v5`

[6] This metric is only applicable to the GSM8K dataset due to its standardized equation notation using «».

duces a systematic approach to enhance diversity within the ISI framework itself.

6 Conclusion

We presented Diversified Iterative Self-Improvement (DIVE), a framework that addresses the challenge of diversity loss in ISI while maintaining model performance. Through systematic experiments on MATH and GSM8k datasets, we demonstrated that our two-component approach – sample pool expansion and data selection – effectively enhances output diversity across multiple evaluation metrics. Our experiments with different sampling sizes and detailed analysis across various difficulty levels demonstrated consistent improvements in diversity without compromising accuracy.

7 Limitations

While our work demonstrates the effectiveness of DIVE in mathematical reasoning tasks, several limitations should be noted:

Task Scope Our study focuses exclusively on mathematical reasoning tasks (MATH and GSM8k). While we evaluate diversity using multiple metrics including equation patterns and embedding-based measures, the generalization of our approach to other domains remains to be explored.

Sampling Strategy Although increasing the sampling size improves diversity, our current approach of fixed sampling per question may not be optimal. Questions of different difficulty levels might benefit from adaptive sampling strategies to better balance computational cost and diversity gains.

Computational Cost Our experiments show that larger sample pools can enhance diversity, but the computational resources required increase significantly with sample size. While our global data usage method provides an efficient alternative to increased sampling, finding the optimal balance between pool size and computational cost remains a challenge.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. 2012. Diverse m-best solutions in markov random fields. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 1–16. Springer.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

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

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. A tale of tails: Model collapse as a change of scaling laws. *Preprint*, arXiv:2402.07043.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Jonathan St BT Evans. 2010. Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, 21(4):313–326.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *Preprint*, arXiv:2404.01413.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

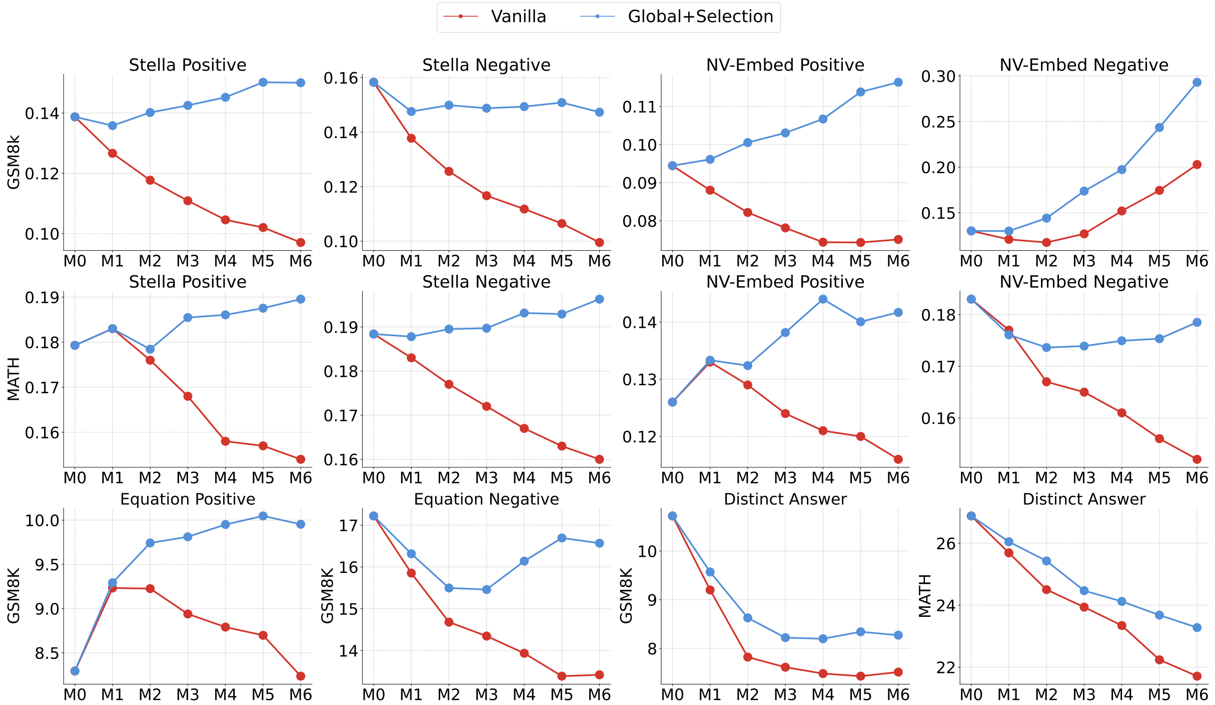

Figure 6: Results of all iterations across different diversity metrics for both the GSM8k and MATH datasets.

de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. Making large language models better reasoners with step-aware verifier. *arXiv preprint arXiv:2206.02336*.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

R Naik, V Chandrasekaran, M Yuksekgonul, H Palangi, and B Nushi. 2023. Diversity of thought improves reasoning abilities of llms. *arXiv preprint*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *CoRR*, abs/2404.19733.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The curse of recursion: Training on generated data makes models forget. *Preprint*, arXiv:2305.17493.

Keith E Stanovich. 2012. On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. *The Oxford handbook of thinking and reasoning*, pages 343–365.

Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. *arXiv preprint arXiv:2004.02990*.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ting Wu, Xuefeng Li, and Pengfei Liu. 2024. Progress or regress? self-improvement reversal in post-training. *CoRR*, abs/2407.05013.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2024. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*.

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, et al. 2024. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *arXiv preprint arXiv:2404.02893*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. 2024. Flow of reasoning: Efficient training of llm policy with divergent thinking. *arXiv preprint arXiv:2406.05673*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Iterative Diversity Results by Alternative Diversity Metrics

Figure 6 shows the full results of all iterations comparing the diversity of "Vanilla" and "Global+Selection" methods across six different diversity metrics, complementing the analysis in Section 4.3. As seen, "Global+Selection" demonstrates higher diversity than "Vanilla" across all iterations and metrics. Moreover, the discrepancy increases with more iterations, highlighting the effectiveness of our method, particularly as the iterative process progresses.