Diverse, not Short: A Length-Controlled Data Selection Strategy for Improving Response Diversity of Language Models

Vijeta Deshpande[1] **Debasmita Ghose**[2] **John D. Patterson**[3]
Roger Beaty[3] **Anna Rumshisky**[1,4]
[1]University of Massachusetts Lowell [2]Yale University
[3]Pennsylvania State University [4]Amazon AGI
vijeta_deshpande@student.uml.edu

Abstract

Diverse language model responses are crucial for creative generation, open-ended tasks, and self-improvement training. We show that common diversity metrics, and even reward models used for preference optimization, systematically bias models toward shorter outputs, limiting expressiveness. To address this, we introduce Diverse, not Short (Diverse-NS), a length-controlled data selection strategy that improves response diversity while maintaining length parity. By generating and filtering preference data that balances diversity, quality, and length, Diverse-NS enables effective training using only 3,000 preference pairs. Applied to LLaMA-3.1-8B and the Olmo-2 family, Diverse-NS substantially enhances lexical and semantic diversity. We show consistent improvement in diversity with minor reduction or gains in response quality on four creative generation tasks: Divergent Associations, Persona Generation, Alternate Uses, and Creative Writing. Surprisingly, experiments with the Olmo-2 model family (7B, and 13B) show that smaller models like Olmo-2-7B can serve as effective "diversity teachers" for larger models. By explicitly addressing length bias, our method efficiently pushes models toward more diverse and expressive outputs[1].

1 Introduction

Alignment has played a key role in making large language models (LLMs) broadly useful, controllable, and safe for real-world applications (Schulman et al., 2017; Bai et al., 2022; Dai et al., 2023; Ouyang et al., 2022; Longpre et al., 2023). As a form of post-training, it typically involves a combination of instruction tuning (Longpre et al., 2023; Peng et al., 2023; Ouyang et al., 2022) and preference optimization (Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023), enabling models to follow human instructions and generate responses that are helpful, harmless, and honest (Bai et al., 2022; Dai et al., 2023). However, alignment comes at a cost: several studies have found that alignment can significantly reduce the diversity of model outputs (Kirk et al., 2023; Doshi and Hauser, 2024; Padmakumar and He, 2023; Anderson et al., 2024; Shaib et al., 2024b; West and Potts, 2025).

This decrease in diversity has important consequences. When humans collaborate with aligned models, the content they produce tends to be less original and less varied (Doshi and Hauser, 2024; Padmakumar and He, 2023). At scale, this reduction in diversity can hinder creative ideation and increase output homogeneity (Anderson et al., 2024; Xu et al., 2024). Beyond creativity, reduced diversity of generated text has a direct impact on the continued improvement of LLMs, with only limited benefits of reduced diversity (Deshpande et al., 2023; Muckatira et al., 2024). Recent studies have shown that repeatedly training models on their own aligned outputs can lead to a consistent decline in diversity, eventually resulting in model collapse (Shumailov et al., 2023; Guo et al., 2023; Seddik et al., 2024).

Despite these challenges, alignment remains essential. The question, then, is not whether to align, but how to preserve or recover the output diversity of aligned models. In this work, we ask: *Can we increase the response diversity of aligned models while retaining the the response quality?*

Prior work has explored a range of strategies to improve output diversity of aligned language models, including methods based on prompting, sampling, and targeted training procedures (Lu et al., 2024; Zhang et al., 2020; Tian et al., 2023; Li et al., 2024, 2025; Lanchantin et al., 2025; Chung et al., 2025; Qin et al., 2025). Sampling techniques such as temperature, top-p, and top-k have been shown to increase diversity, though often at the cost of

[1]Code and data can be accessed at,
Code: Diverse-NS-Repo,
Dataset: Diverse-NS-Preference-Learning-Data

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33906–33927
November 4-9, 2025

reduced quality (Zhang et al., 2020). Sequential prompting strategies are also helpful in increasing response diversity (Lu et al., 2024; Tian et al., 2023). However, the computational cost scales rapidly with more discussion turns due to increasing context length. Training approaches have introduced explicit diversity objectives (Li et al., 2025; Chung et al., 2025; Cideron et al., 2024) and entropy regularization (Li et al., 2024) to encourage more varied outputs. Self-learning methods, where the model generates its own training data, have also been used to promote diversity (Tian et al., 2024; Lanchantin et al., 2025; Qin et al., 2025).

However, one critical confound, text length, has received little scrutiny in recent work. Widely used diversity metrics are length-sensitive and consistently assign higher scores to shorter passages (Covington and McFall, 2010; McCarthy and Jarvis, 2010; Shaib et al., 2024a). While this bias is less problematic in structured generation tasks, optimizing these metrics can reduce expressiveness in open-ended writing, which thrives on depth and nuance, thereby undermining the very creativity they are meant to cultivate. But even though optimizing length-sensitive metrics can clearly backfire, the role of length in both measuring and improving diversity has been largely overlooked. Our work aims to close this gap.

To address this overlooked confounding factor, we propose *Diverse, not Short* (Diverse-NS), a length-controlled data selection strategy that counteracts the hidden brevity bias in standard diversity metrics and improves diversity in both structured and free-form generation. The framework first uses sequential prompting to elicit more diverse responses, followed by preference pair curation that improve both diversity and quality while maintaining comparable response lengths (within ± 5 words). Using these preference pairs, we apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) to improve the response diversity of the base model. Our key contributions are:

1. **Diverse-NS:** A length-controlled data selection strategy that significantly improves the response diversity of Llama-3.1-8B and Olmo-2-7B using only 3k preference pairs.
2. **Diverse-NS-Lite:** A computationally efficient variant that achieves comparable performance to Diverse-NS while significantly reducing the data filtering cost.
3. **Small-to-large transfer:** We highlight the potential of smaller models to serve as effective "diversity teachers" for larger variants, enabling low-cost diversity alignment.
4. **Length-controlled diversity evaluation:** We introduce *Diversity Decile*, a diversity metric that adjusts for text length.
5. **Dataset:** We release a high-quality dataset of 6k preference pairs generated from Llama-3.1-8B and Olmo-2-7B to support future research on length-aware diversity alignment.

2 Related Work

Increasing Diversity without Training. Zhang et al. (2020); Chung et al. (2023), shows that common sampling methods such as temperature, top-p, top-k, are comparable in terms of increasing the diversity but, increasing diversity often comes at the price of reduced quality. For curating a generic large-scale dataset, prompting methods can boost topical, stylistic, and formatting diversity (Li et al., 2023; Chen et al., 2024; Face, 2024; Ge et al., 2024). Conversely, for more task-specific datasets, sequential prompting can elicit diverse responses (Lu et al., 2024; Tian et al., 2023; Qin et al., 2025).

Increasing Diversity with Training. Augmenting method-specific objective functions with elements that directly maximize diversity has been successful in increasing response diversity (Li et al., 2024; Chung et al., 2025; Li et al., 2015, 2025). The other approach gaining more attention in recent studies is to adopt a three-step procedure: generate diverse data, filter data for improving quality, and fine-tune LLM on the filtered data (Lanchantin et al., 2025; Chung et al., 2025; Qin et al., 2025). This approach has been successful in task-specific alignment, but more generic self-training has still seen limited success (Li et al., 2023; Face, 2024; Shumailov et al., 2023; Guo et al., 2023; Herel and Mikolov, 2024; Seddik et al., 2024). Our work is closest to the task-specific alignment studies in the self-learning framework (Lanchantin et al., 2025; Qin et al., 2025).

Diversity Evaluation. Evaluation of diversity is challenging for a few reasons: length bias (McCarthy and Jarvis, 2010; Covington and McFall, 2010; Mass, 1972; Johnson et al., 2023; Deshpande et al., 2025), relative difficulty in achieving substantial agreement between humans (Chakrabarty et al., 2023, 2024; Gómez-Rodríguez and Williams, 2023), and inconsistent human preferences (Evans

et al., 2016). Despite the challenges, many studies have highlighted the compromised diversity of synthetic or human-LLM collaborative text (Shaib et al., 2024b,a; Salkar et al., 2022; Padmakumar and He, 2023; Guo et al., 2023; Kirk et al., 2023; Doshi and Hauser, 2024; Anderson et al., 2024). So, we present *Diverse-NS*, to increase the response diversity and propose a metric, *Diversity Decile*, to measure diversity in a length-controlled way.

3 Preliminaries

Self-learning, also known as self-training, is a semi-supervised approach involving three main steps: data generation (pseudo-labeling), data filtering, and model learning (Lee et al., 2013; Amini et al., 2025). In our setup, data generation involves sampling text from a language model in response to story-writing prompts. This is followed by filtering, where we construct high-quality preference pairs—two continuations for the same prompt, with one preferred over the other. We refer to the preferred continuation as the "chosen" and the other as the "rejected". Using this preference dataset, we apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) to train the model to favor the chosen responses.

4 Data

We describe data generation and filtering pipeline designed to elicit diverse model responses for downstream preference tuning. The pipeline first generates candidate stories using a sequential prompting strategy, then filters the pool of generated responses to form preference pairs suitable for Direct Preference Optimization (DPO) training (Rafailov et al., 2023). The preference pairs are formed to maximize the diversity and quality gain while maintaining the same length for "chosen" and "rejected" samples.

4.1 Data Generation

Task Setup. We focus on a creative writing task to build the dataset for preference learning. The goal is to generate short stories (five sentences) that must include three words specified in the prompt. This task has been extensively validated in studies of human creativity (Prabhakaran et al., 2014). To create a diverse set of prompts, we first curated a list of 300 unique words, W_u[2]. For generating short stories from LMs, we create prompts by randomly sampling three-word sets from W_u.

Sequential Prompting. Given the task setup, we create 1k story writing prompts, with 1k unique three-word sets. The exact prompt is provided in Appendix A.1. We initially sampled 10k stories (10 per prompt) using a temperature of 1.0 from each of the following LMs: Llama-8B and Olmo-7B (Grattafiori et al., 2024; OLMo et al., 2024). Within the sampled stories, we extracted the repeating Part-Of-Speech (POS) bigrams and found that the start of the story is highly likely to have repetitions across different prompts (refer to Table B.1). To overcome these repetitions, we performed a second inference call to re-draft the story with additional constraints, an approach similar to *Denial Prompting* presented by Lu et al. (2024) (refer to Appendix A.1 exact prompt). In our case, unlike Lu et al. (2024), the constraints we use are specifically targeted to elicit a more diverse response from the model while maintaining the same (or comparable) length. With a pilot analysis on the initial 20k responses, we find that the story generated in the second inference call is on average more diverse (refer to Table B.2). These results motivated us to set up the final two-step data generation process, first inference call to collect natural responses from the model, and second inference call to redraft the natural response into a more diverse story. In the final data generation phase, we used 20k unique three-word sets to generate prompts and sampled 10 first and second responses for each prompt, resulting in a dataset of 200,000 tuples of prompt, first response, and second response, per model (Llama-8B and Olmo-7B). We denote the data as follows: $\mathcal{D}^{(\pi)} = \{(p, r_1, r_2)_i \mid i = 1, \ldots, 200{,}000\}$ where, p, r_1, and r_2 denote the prompt, first response, and second response, respectively, generated from model (policy) π. Note that $|\{p_1, p_2, \ldots, p_{200{,}000}\}| = 20,000$ and we use two models, $m \in \{\text{Llama-8B, Olmo-7B}\}$, for data generation.

4.2 Data Filtration

The Chosen and Rejected Pools. Each instance in our generated dataset is a tuple (p, r_1, r_2), where p is the prompt and r_1, r_2 are two responses conditioned on it. The first response r_1 reflects the model's default behavior which are stories generated without intervention, capturing its most likely completion. In contrast, the second response r_2

[2]A manually curated list of 20 words was extended using GPT-4o and Claude-3.7.

is generated with additional instructions aimed at reducing repetition, resulting in a more diverse output. We leverage this contrast by designating r_1 as the *rejected* response and r_2 as the *chosen* one. This setup encourages the model to prefer more diverse continuations that it is already capable of generating. Hence, it provides a strong self-learning framework for improving diversity.

Filtration Rules. Each pair (r_1, r_2) gives us a natural candidate for rejected and chosen responses. On average, the second response r_2 is more diverse than the first r_1 (Table B.2), but not every pair guarantees learning higher diversity. To ensure that the model receives consistent and useful learning signals, we apply a set of filtering rules.

First, we require that the diversity of r_2 exceeds that of r_1, so that the model consistently learns to prefer more diverse continuations. However, higher diversity may negatively impact text quality as prior work has shown a trade-off between the two (Zhang et al., 2020). To ensure that preference learning also promotes higher quality, we further require that r_2 be of higher quality than r_1. Additionally, we filter out cases where both r_1 and r_2 are of poor quality, even if r_2 is marginally better. To do so, we enforce that r_2 must surpass the median quality of all r_1 responses. Lastly, most diversity metrics have been shown to be negatively correlated with text length (Covington and McFall, 2010; Shaib et al., 2024a; McCarthy and Jarvis, 2010), which introduces a bias toward shorter texts. This issue has not been explicitly addressed in the recent studies for training and evaluation of LMs for diversity (Qin et al., 2025; Lanchantin et al., 2025; Chung et al., 2025). To control for this, we constrain r_1 and r_2 to be of comparable length (± 5 words). Ideally, we would like r_1 and r_2 to have exactly the same length. However, in practice, very few examples satisfy this strict constraint, especially when working with smaller language models (under 10B parameters). Therefore, we relax the constraint and allow a maximum length difference of ± 5 words between r_1 and r_2.

In summary, we retain a data point for preference learning only if it satisfies all of the following conditions, applied in order:

- The quality of r_2 is greater than or equal to the 50^{th} percentile of all r_1 quality scores.
- The quality of r_2 is greater than r_1.
- The diversity of r_2 is greater than r_1.
- The absolute difference in word count between r_1 and r_2 is at most five words.

Diversity and Quality Metrics. We use entropy to measure diversity and the ArmoRM reward model scores (Wang et al., 2024) to assess quality[3]. Entropy is a standard metric for lexical diversity (Lanchantin et al., 2025), with higher values indicating greater diversity. In our self-learning setup, entropy is useful because it reflects the model's likelihood of producing a certain continuation of the prompt. When used in filtering, it helps identify training data that aligns with the model's own capabilities. For each example, we compute the entropy and the reward model score of both r_1 and r_2, conditioned on the original prompt p. When we use our data generation method, and use entropy and ArmoRM values for filtration, we call our approach, Diverse, not Short (Diverse-NS or D-NS).

Lightweight Filtration. While entropy and ArmoRM scores are high-quality metrics for measuring diversity and response quality, they are computationally expensive. Each example (p, r_1, r_2) requires two additional inference calls to compute entropy and two more for ArmoRM scoring. To reduce this overhead, we evaluated seven alternative metrics and measured their correlation with entropy and ArmoRM scores. Among these, Type-Token Ratio (TTR) showed the highest correlation with entropy (Pearson $r = 0.2027$, $p < 0.0001$), and the MAAS index (Mass, 1972) was most correlated with ArmoRM scores (Pearson $r = 0.2357$, $p < 0.0001$). Refer to Table 1 for all correlation results. Based on these findings, we replace entropy with TTR and ArmoRM scores with MAAS in our filtering pipeline. When this lightweight variant is used during data filtering, we refer to the resulting method as Diverse-NS-Lite (or D-NS-Lite). We provide a discussion on the computational savings achieved with Diverse-NS-Lite in Appendix D.

Post-Filtration Properties. Based on the correlation analysis (Table 1), it is worth noting that both entropy and ArmoRM scores are negatively correlated with text length. As a result, optimizing for diversity or quality alone may unintentionally favor shorter responses as the "chosen" continuations. To avoid this bias, it is essential to explicitly control for length when curating preference learn-

[3] refer to Appendix C for a brief description of ArmoRM scores

Method	Word Count	TTR	MATTR	HD-D	MTLD	MAAS
Entropy	−0.1574	**0.2027**	0.0800	0.1071	0.0656	−0.1104
ArmoRM Score	−0.3461	0.1698	−0.0042**	−0.0487	0.0749	**0.2357**

Table 1: **Correlation Analysis.** Pearson correlation coefficients between six text statistics and two target metrics: entropy (diversity) and ArmoRM reward scores (quality). Both entropy and ArmoRM scores show negative correlation with text length. Among diversity metrics, TTR exhibits the strongest correlation with entropy, while the MAAS index shows the highest correlation with ArmoRM scores. **: $p < 0.001$; all others: $p < 0.0001$.

Method	Num. Pref. Pairs	Word Count Δ
No Filtering	200,000	-0.68 ± 11.33
DivPO	3,000	-49.90 ± 17.51
Ours - D-NS-Lite	3,000	-0.90 ± 2.91
Ours - D-NS	3,000	-1.35 ± 2.93

Table 2: **Data Properties After Filtering.** This table reports the average (±std.dev.) length difference (Δ) between *chosen* and *rejected*. While DivPO tends to favor significantly shorter *chosen* responses.

ing data for improving diversity. To show this, we implement a recent study that is closest to our method, Diverse Preference Optimization(DivPO) (Lanchantin et al., 2025) (refer to Appendix E for implementation details). DivPO also generates responses and filters the responses to form preference learning pairs without explicitly controlling the length of the chosen and rejected continuations. We compare pre- and post-filtration data properties for DivPO and Diverse-NS in Tab. 2. The table clearly shows that in the pursuit of maximizing the entropy values, DivPO selects significantly shorter (-49.90 words shorter on average) responses as the *chosen* responses in the final preference data.

5 Experimental and Evaluation Setup

5.1 Preference Tuning

After generating and filtering the data, we fine-tune the same base policy π that was used to generate it. In other words, data generated by Llama-8B is used to train Llama-8B, and likewise for Olmo-7B. To ensure a fair comparison across methods (DivPO, D-NS, and D-NS-Lite), we limit the final training dataset to 3,000 preference pairs[4]. To construct this 3k dataset, we first compute the entropy gain for each pair as the difference between the entropy of the *chosen* and *rejected* responses [5]. We then sort all pairs by entropy gain in descending order and select the top 3k examples. This ensures that the final training set maximizes diversity gain for the base model. The same selection procedure is applied to all three methods.

We further extend our experiments to evaluate the utility of training larger models with data generated from smaller ones. For this, we train Olmo-13B using preference pairs generated from Olmo-7B. We provide all hyperparameter values in Appendix F. All experiments are run on a single NVIDIA RTX 6000 GPU (48GB memory), using a per-device batch size of 2 and a global batch size of 64. Training Llama-8B or Olmo-7B takes approximately 100–150 minutes while O-13B takes 200-220 minutes per run, highlighting our setup efficiency.

5.2 Evaluation

5.2.1 Tasks

We evaluate the model's response diversity with four tasks: Divergent Association Task (DAT), Persona Generation Task (PGT), Alternate Uses Task (AUT), and Creative Writing Task (CWT).

Divergent Associations Task (DAT). The DAT (Olson et al., 2021) is a psychological test commonly used to assess divergent thinking in humans. Participants are asked to generate a list of 10 words that are as dissimilar from each other as possible. Recent studies have adapted DAT to evaluate the creativity of language models, focusing on their ability to produce diverse outputs (Bellemare-Pepin et al., 2024). To quantify model performance on DAT, we use the Divergent Semantic Integration (DSI) metric (Johnson et al., 2023), which computes the average semantic distance of each word in the generated list from all others. Higher DSI values indicate more divergent thinking and greater ideological diversity. Following Johnson et al. (2023), we extract token embeddings from the 6^{th} layer of BERT-large for the generated list and compute the average pairwise cosine distance between all embeddings. This approach has been

[4]We observed that the size of the dataset after filtering is the smallest for Diverse-NS, slightly more than 3k. Hence, to make the training runs more comparable across methods, we limit the size of the dataset to 3k for all methods.

[5]note that, by construction, the *chosen* response has higher entropy in the filtered set

shown to correlate strongly with human judgments of creativity (Johnson et al., 2023). We provide the exact prompt used for DAT in Appendix A.2. For a robust evaluation, we sample 100 DAT responses per model using temperature 1.0 and different random seeds. From these 100 lists (each with 10 words), we compute and report two metrics: (1) the average and standard deviation of DSI scores, and (2) the number of unique words across all 1,000 generated tokens. In both cases, higher values indicate greater diversity.

Persona Generation Task (PGT). To assess diversity in structured generation, we use the PGT, also used in the study conducted by Lanchantin et al. (2025). In this task, the model is prompted to generate a JSON object with three fields: first name, city of birth, and current occupation to evaluate the model's ability to produce varied persona descriptions. The exact prompt is provided in Appendix A.2. We sample 100 responses per model using temperature 1.0 and different random seeds. For each key in the JSON object, we report the proportion of unique values across the 100 responses. Higher uniqueness indicates greater diversity.

Alternate Uses Task (AUT). The Alternate Uses Task (AUT) is a common and rigorously validated psychological test to measure human divergent thinking (Guilford, 1956). In this task, the subject/model is asked to generate creative and unconventional uses for objects (e.g., broom). The prompt and list of objects used for evaluation are provided in Appendix A.2. We use 15 unique objects and generate 10 responses per object using different random seeds, resulting in 150 total responses sampled at temperature 1.0. For quantifying the diversity of the generated uses, we measure the distance between the target object and generated uses with the help of BERT-large encodings, a validated approach that correlates with human creativity ratings (Patterson et al., 2023). We report the mean and standard deviation of the distance values, higher values indicate higher diversity.

Creative Writing Task (CWT). The CWT — based on a well-validated psychological assessment of creativity (Prabhakaran et al., 2014) — is exactly the same as our data generation task. That is, given a set of three words, the subject/model is tasked with generating a creative short story that includes all three words. We provide a separate list of three-word sets used for evaluation in Appendix A.2. We sample 10 responses for each of the seven three-word sets with temperature of 1.0. Unlike our other evaluation tasks, we measure the diversity as well as the quality of the generated responses. Similar to Johnson et al. (2023), we calculate the DSI metric to measure the diversity of the generated story. For quality measurements, we resort to the ArmoRM reward model preference scores (Wang et al., 2024). We report DSI, ArmoRM, and 4-gram diversity values, where higher values are more desirable for all metrics.

5.2.2 Length-Adjusted Evaluation

While most diversity metrics exhibit bias toward shorter outputs, Johnson et al. (2023) shows that the DSI metric displays the opposite tendency-it favors longer responses. This is not an issue in tasks like DAT, where the output length is fixed at 10 words. But for open-ended tasks such as CWT, longer stories may receive disproportionately high DSI scores primarily due to their length, rather than genuine diversity. To address this issue, we introduce a novel evaluation metric: Δ *Diversity Decile (Δ DD)*, which accounts for text length when assessing diversity.

Change in Diversity Decile (ΔDD). We first build a *decile map* that captures the empirical distribution of diversity scores at each length. Using 800 000 stories collected from Llama-8B and Olmo-7B over 40 000 prompts, we: (1) group responses by word count w; (2) compute decile thresholds for a chosen diversity metric (e.g. TTR, MTLD); and (3) store these percentile thresholds in a lookup table $\mathcal{M}$. Here, a *decile* refers to one of ten intervals that divide the distribution of diversity scores for a given length into ten equal parts. The top decile corresponds to the most diverse 10% of responses at that length, the second-highest to the next 10%, and so on. This mapping allows us to estimate the *approximate diversity rank* of any new response relative to other responses of the same length. At evaluation time, a new response r with word count w_r and diversity score d_r is assigned the highest decile index $k \in \{0, \ldots, 9\}$ such that d_r exceeds the k-th threshold in $\mathcal{M}[w_r]$. Formally, $DD(r, \mathcal{M}) = k$, where larger k means the response is more diverse than a greater share of previously observed texts of the *same length*[6].

[6]If the response length of the new response is not present in the constructed map, we select the closest neighbor as an approximation. However, it is important to note that due to the ample size of our synthetic data, we never encounter a case

Task	Metric	Base Model	DivPO	Ours D-NS-Lite	Ours D-NS
LLaMA-8B					
DAT	DSI	0.7535	0.7545	0.7590	**0.7640**
DAT	Unique Words	0.4575	0.4593	0.4797	**0.4914**
PGT	Unique First Names	0.6500	0.6100	**0.6900**	**0.6900**
PGT	Unique Cities	0.3300	0.3100	**0.4700**	0.4200
PGT	Unique Occupations	0.4100	0.3900	**0.5100**	0.4900
AUT	DSI	0.8876	0.8837	0.8876	**0.8878**
CWT	DSI	0.8515	0.8521	0.8556	**0.8581**
CWT	ArmoRM Score	0.1451	**0.1495**	0.1369	0.1405
CWT	4-gram div.	2.8550	2.9320	2.9450	**2.9620**
OLMo-7B					
DAT	DSI	0.7480	0.7509	**0.7662**	0.7639
DAT	Unique Words	0.6139	0.6079	**0.6347**	0.6327
PGT	Unique First Names	0.3300	0.3300	0.3300	**0.3400**
PGT	Unique Cities	**0.3100**	0.3000	0.2700	0.2700
PGT	Unique Occupations	0.5200	0.5500	**0.6100**	**0.6100**
AUT	DSI	0.8836	0.8846	0.8852	**0.8858**
CWT	DSI	0.8499	0.8491	0.8548	**0.8563**
CWT	ArmoRM Score	0.1435	0.1441	0.1462	**0.1464**
CWT	4-gram div.	3.1270	3.1690	**3.1750**	3.1620
OLMo-13B					
DAT	DSI	0.7233	0.7282	0.7320	**0.7364**
DAT	Unique Words	**0.3421**	0.3340	0.3310	0.3256
PGT	Unique First Names	0.4100	0.4100	0.4400	**0.4500**
PGT	Unique Cities	0.3500	0.3500	0.3700	**0.3900**
PGT	Unique Occupations	0.1900	0.1900	0.1900	**0.2000**
AUT	DSI	0.8943	0.8960	**0.8974**	0.8970
CWT	DSI	0.8557	0.8555	**0.8616**	0.8614
CWT	ArmoRM Score	0.1571	0.1589	0.1585	**0.1590**
CWT	4-gram div.	3.0820	3.0770	3.095	**3.1070**

Table 3: **Diversity and Quality Evaluation.** We present the average diversity (DSI or unique values) and quality (ArmoRM Score) measurements for model responses collected on four creative generation tasks (Structured Gen.: DAT, PGT, Free-Form Gen.: AUT, CWT).

To evaluate the effect of preference tuning, we average DD scores over 70 CWT prompts for the base and the preference-tuned models and report their difference: $\Delta DD = \overline{DD}_{\text{tuned}} - \overline{DD}_{\text{base}}$.

Positive ΔDD values indicate improved diversity, with higher values corresponding to a larger improvement. Negative values signify reduced diversity, and $\Delta DD = 0$ signifies no change. Note that, DD is agnostic to the choice of diversity metric. We therefore report ΔDD values using seven standard metrics: TTR, MATTR, HD-D, MTLD, and MAAS. We also compute ΔDD using ArmoRM reward scores to quantify the gain or loss in quality. This length-aware normalization prevents either long or short responses from being over-credited for diversity[7].

where the response length value is not found in the constructed map.

[7]We provide a summary of all metrics in Table J.1

6 Results

Divergent Associations Task (DAT). In our DAT evaluation (Tab. 3), we see that both Diverse-NS and its lightweight variant deliver clear improvements in diversity over the untrained base and the DivPO baseline across all model sizes. Remarkably, even the D-NS-Lite variant consistently outperforms DivPO, demonstrating that a compact diversity strategy can be highly effective. Interestingly, using data generated by the smaller Olmo-7B to fine-tune the larger Olmo-13B yields diversity gains for every method, highlighting how smaller models can serve as powerful "diversity teachers" for their larger counterparts.

Persona Generation Task (PGT). In our PGT evaluation (Tab. 3), Diverse-NS produces more distinct first names, cities, and occupations than DivPO for every model, with the sole exception of the city metric on Olmo-7B. Outside that one case,

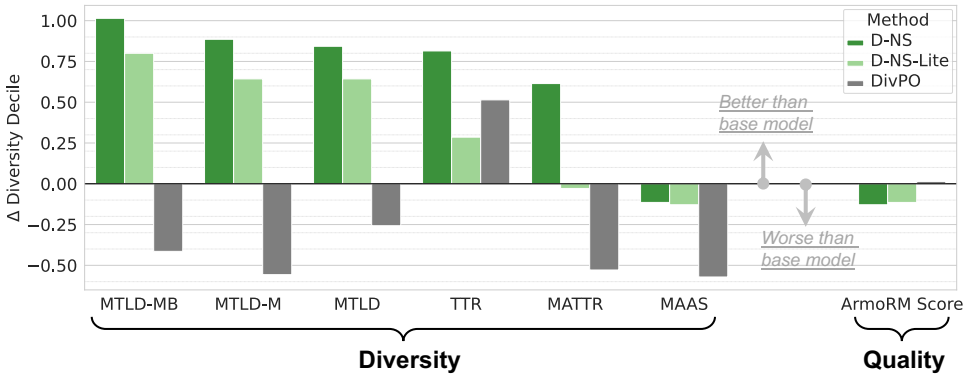

Figure 1: **Diversity and Quality Evaluation on CWT.** This figure shows ΔDiversity Decile (ΔDD) values (y-axis) across various metrics (x-axis), computed from 70 CWT responses generated by the Olmo-2-7B model. A value of zero represents base model performance; bars indicate improvements from preference-tuned models. *D-NS* achieves the highest diversity gains overall, while *D-NS-Lite* consistently outperforms *DivPO*, except under TTR. In terms of quality (ArmoRM), *DivPO* shows a slight improvement, whereas our methods show a minor drop.

Diverse-NS-Lite also outperforms DivPO across all three metrics. Notably, on Llama-8B, Diverse-NS-Lite matches or exceeds the baseline and Diverse-NS on every attribute of the task.

Alternate Uses Task (AUT). In our AUT evaluation (Tab. 3), Diverse-NS-Lite consistently beats DivPO, and Diverse-NS consistently beats Diverse-NS-Lite, though only by a small margin.

Creative Writing Task (CWT). In our CWT evaluations (Tab. 3), Diverse-NS produces the highest DSI scores for both Llama-8B and Olmo-7B. Interestingly, for Llama-8B the other methods actually reduce the ArmoRM score below baseline but Diverse-NS exceeds it. The highest 4-gram diversity is observed for Diverse-NS or -Lite in all cases. We also compute ΔDD with six lexical diversity measures and ArmoRM. Both Diverse-NS and its lightweight variant significantly outperform DivPO on every diversity metric. The ΔDD remains above the baseline for all metrics except MAAS, where it dips marginally below and similarly shows a slight under-performance for ArmoRM. Crucially, even where ΔDD suggests a minor quality drop, the absolute diversity values after self-training still exceed those of the base model (despite longer outputs), indicating that any loss in writing quality is minimal (refer to Appendix I for Llama-8B and Olmo-13B results)[8].

Significance of DSI Improvements. In line with the findings of Johnson et al. (2023), we observe a narrow distribution of DSI values in our CWT experiments (percentiles: $5^{th}, 95^{th}$: $0.84, 0.87$). Given the narrow distribution, to highlight the significance of the DSI improvements achieved with D-NS or D-NS-Lite, we conduct an independent t-test on method pairs, e.g., base model versus D-NS, with 70 DSI measurements (for 70 unique CWT responses) for each method. Results are presented in Table 4, where negative t-statistics indicate that "Method B" yields higher response diversity than "Method A". We find that both D-NS and D-NS-Lite produce significantly more diverse responses than the base model and DivPO. On the other hand, the differences between the base model and DivPO, as well as between D-NS and D-NS-Lite, are not statistically significant. These findings further underscore the effectiveness of D-NS-Lite as a computationally efficient alternative to D-NS.

Human Evaluation. To further understand the human perceived value of DSI improvements, we conduct a human evaluation on the CWT responses. Under pairwise evaluation, annotators are shown a prompt along with two responses from two different methods (e.g., D-NS versus DivPO) and asked to select the more diverse response (see Appendix A.2). Per CWT prompt, we select 20 least similar (Jaccard sim.) pairs to ensure maximally distinguishable comparisons between methods. Using 7 prompts, we construct 140 pairs each for D-NS vs. Base and D-NS vs. DivPO. The resulting 280 comparisons are evenly and randomly assigned to four annotators: two involved in the study (E1, E2) and two independent evaluators with no prior exposure (E3, E4). Human evaluation confirms that D-NS consistently outperforms both baselines in win percentage. Except for E3, all annotators

[8]We provide all results with std. dev. values in Table H.1

Method A	Method B	t-statistic	p-value
LLaMA-8B			
Base	DivPO	-0.4126	Not Significant
Base	D-NS	-4.1208	< 0.001
Base	D-NS-Lite	-2.6100	< 0.05
DivPO	D-NS	-3.7516	< 0.001
DivPO	D-NS-Lite	-2.2214	< 0.05
D-NS	D-NS-Lite	1.5492	Not Significant
Olmo-7B			
Base	DivPO	0.3609	Not Significant
Base	D-NS	-3.4402	< 0.001
Base	D-NS-Lite	-2.7348	< 0.01
DivPO	D-NS	-3.5529	< 0.001
DivPO	D-NS-Lite	-2.8998	< 0.01
D-NS	D-NS-Lite	0.9972	Not Significant

Table 4: **Significance of DSI Improvements.** In this table, we present the results of independent t-tests on DSI scores between pairs of methods. Negative t-values indicate that "Method B" yields more diverse responses than "Method A". We highlight cases (in green) where our proposed methods (D-NS or D-NS-Lite) are significantly better than the base model or DivPO.

strongly preferred D-NS, with win rates reaching as high as 72.73% against DivPO (E4). These findings suggest that even modest improvements in DSI scores correspond to clear and consistent gains in human-perceived diversity. We extended our human evaluation to the LLM-as-a-Judge evaluation that covers comparison between D-NS-Lite and the baselines as well. The results of the LLM-as-a-Judge experiment reciprocate the human evaluation findings and underscore the effectiveness of both D-NS and D-NS-Lite (refer to Appendix K for details). We provide a few examples of the model responses in Appendix L for the readers.

7 Discussion

We introduced *Diverse-NS*, a data selection strategy to improve output diversity while preserving quality. Experiments with Llama-8B and Olmo-7B show that *Diverse-NS* improves diversity on four creative generation tasks: DAT, PGT, AUT, CWT. On CWT, the diversity gains achieved by *Diverse-NS* and its lightweight variant, *Diverse-NS-Lite*, are statistically significant. These improvements are further supported by human evaluations, where *Diverse-NS* achieves higher win rates against both the base model and DivPO.

***Diverse-NS* is highly efficient.** All gains are achieved with only 3k preference pairs and less than two hours of training on a single 48 GB GPU. The lightweight variant, *Diverse-NS-Lite*, replaces

Human Evaluator	D-NS vs. Base (D-NS Win% - Tie%)	D-NS vs. DivPO (D-NS Win% - Tie%)
E1	$63.16 - 28.95$	$69.70 - 24.24$
E2	$72.22 - 0.00$	$63.64 - 0.00$
E3	$50.00 - 5.88$	$51.52 - 3.03$
E4	$64.71 - 11.76$	$72.73 - 15.15$

Table 5: **Human Evaluation on CWT.** Pairwise human evaluation results comparing D-NS against the base model and DivPO on the CWT task. Each cell shows the win and tie percentages for D-NS. D-NS consistently achieves higher win rates against both baselines.

costly entropy and ArmoRM scoring with inexpensive proxies yet still surpasses DivPO in nearly every setting. We further show that a 7B model can act as an effective "diversity teacher" for its 13B counterpart, pointing to a low-cost path for diversity-aware alignment at scale.

***Diverse-NS* maintains high quality.** Diversity and quality are often at odds (Zhang et al., 2020; Chung et al., 2023), and we observe this trade-off in our experiments as well. However, there are encouraging instances where both improve together. For Olmo-7B and Olmo-13B, the ArmoRM score increases alongside diversity. ΔDiversity Decile values further confirms that, for Olmo-13B, diversity and quality consistently rise in tandem. In other cases, we observe only a minor drop in quality, suggesting that *Diverse-NS* effectively balances this trade-off in most scenarios.

The long-standing challenge of length. Evaluating diversity remains difficult due to the well-known length bias in most diversity metrics. This issue extends to ArmoRM scores, which also favor shorter texts (Tab. 1), further complicating reliable evaluation. To mitigate this, we introduce the Δ*Diversity Decile* metric, which quantifies percentile gains or losses in diversity (or quality) relative to the base model. Using this length-adjusted metric, we observe substantial improvements in diversity across most lexical diversity measures, along with small but mixed changes in quality.

Overall, *Diverse-NS* offers a practical and scalable solution for boosting diversity in aligned LLMs. By addressing the length bias in both training and evaluation, our sets a foundation for more expressive and diverse language generation. We hope this work encourages further exploration of length-aware diversity alignment.

Limitations

While our study demonstrates the effectiveness of diversity-aware self-learning, several areas remain open for future exploration. First, our data filtering relies on a single diversity metric (e.g., entropy or TTR). Although effective, no single metric can fully capture all aspects of text diversity. Future work could incorporate multiple metrics to jointly optimize lexical, semantic, and syntactic variation, as well as novelty, to better capture diverse training signals. Second, we focus on one data generation task—short story writing—which allows for controlled analysis and task-specific improvements. Expanding the framework to include a broader set of tasks could lead to more generalizable diversity enhancements. Third, our self-learning setup investigates only a single round of preference tuning. While this provides a strong baseline, recent work suggests that repeated rounds of self-training can affect diversity (Guo et al., 2023; Seddik et al., 2024; Herel and Mikolov, 2024). It would be valuable to study how diversity evolves across multiple self-learning iterations in our framework. It is worth noting a peculiar change in the length distribution of the preference-tuning model (Table G.1). Even though preference pairs are of comparable lengths in *Diverse-NS* and *Diverse-NS-Lite*, the model learns to be more expressive. We suspect this shift is influenced by a skewed proportion of longer preference pairs, which may inadvertently bias the model toward generating longer responses. Controlling the length distribution is challenging under our current framework due to the strict filtering criteria. In future work, we aim to address this by extending our method to a multi-task setup that includes both short and long generation tasks.

Ethics Statement

Our work focuses on improving the diversity of language model outputs, particularly in creative and open-ended tasks. While diversity is an important dimension of language generation, it may come at the cost of factual correctness in certain scenarios. Therefore, we caution against the use of our dataset or models in tasks where factual accuracy is critical, such as medical advice, legal reasoning, or scientific fact-checking. We also acknowledge the growing computational divide in language model research. A key motivation behind our approach is to make diversity-aware alignment more accessible. By limiting training to 3,000 preference pairs and demonstrating the effectiveness of smaller models (e.g., Olmo-2-7B) as diversity teachers, we aim to lower the resource barrier and encourage further research in compute-constrained environments. Finally, while we use proprietary language models (such as GPT-4o and Claude) to assist in editing and refining text during data curation and paper writing, no portion of this manuscript was generated entirely by an LLM. All content has been written, reviewed, and edited by the authors to ensure clarity, originality, and scientific rigor.

Acknowledgment

This work was funded in part by an Amazon AGI research award to Anna Rumshisky.

References

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. 2025. Self-training: A survey. *Neurocomputing*, 616:128904.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2024. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. *arXiv preprint arXiv:2409.14509*.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. 2024. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*.

Geoffrey Cideron, Andrea Agostinelli, Johan Ferret, Sertan Girgin, Romuald Elie, Olivier Bachem, Sarah Perrin, and Alexandre Ramé. 2024. Diversity-rewarded cfg distillation. *arXiv preprint arXiv:2410.06084*.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Vijeta Deshpande, Ishita Dasgupta, Uttaran Bhattacharya, Somdeb Sarkhel, Saayan Mitra, and Anna Rumshisky. 2025. A penalty goes a long way: Measuring lexical diversity in synthetic texts under prompt-influenced length variations. *arXiv preprint arXiv:2507.15092*.

Vijeta Deshpande, Dan Pechi, Shree Thatte, Vladislav Lialin, and Anna Rumshisky. 2023. Honey, i shrunk the language: Language model behavior at reduced scale. *arXiv preprint arXiv:2305.17266*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Anil R Doshi and Oliver P Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Owain Evans, Andreas Stuhlmüller, and Noah Goodman. 2016. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Hugging Face. 2024. Cosmopedia: An open source mixture of experts for retrieval-augmented generation. https://huggingface.co/blog/cosmopedia. Accessed: 2025-05-16.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*.

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

J. P. Guilford. 1956. The structure of intellect. *Psychological Bulletin*, 53(4):267–293. Place: US Publisher: American Psychological Association.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.

David Herel and Tomas Mikolov. 2024. Collapse of self-trained language models. *arXiv preprint arXiv:2404.02305*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*.

Dan R Johnson, James C Kaufman, Brendan S Baker, John D Patterson, Baptiste Barbot, Adam E Green, Janet van Hell, Evan Kennedy, Grace F Sullivan, Christa L Taylor, et al. 2023. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 55(7):3726–3759.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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*.

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

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

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*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

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

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. 2024. Entropic distribution matching for supervised fine-tuning of llms: Less overfitting and better diversity. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. 2024. Benchmarking language model creativity: A case study on code generation. *arXiv preprint arXiv:2407.09007*.

Heinz-Dieter Mass. 1972. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. 2024. Emergent abilities in reduced-scale generative language models. *arXiv preprint arXiv:2404.02204*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences of the United States of America*, 118(25). Place: United States.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

John D Patterson, Hannah M Merseal, Dan R Johnson, Sergio Agnoli, Matthijs Baas, Brendan S Baker, Baptiste Barbot, Mathias Benedek, Khatereh Borhani, Qunlin Chen, et al. 2023. Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts*, 17(4):495.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Ranjani Prabhakaran, Adam E. Green, and Jeremy R. Gray. 2014. Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods*, 46(3):641–659. Place: Germany Publisher: Springer.

Yiwei Qin, Yixiu Liu, and Pengfei Liu. 2025. Dive: Diversified iterative self-improvement. *arXiv preprint arXiv:2501.00747*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nikita Salkar, Thomas Trikalinos, Byron C Wallace, and Ani Nenkova. 2022. Self-repetition in abstractive neural summarizers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2022, page 341.

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

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. 2024. How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. 2024a. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024b. Detection and measurement of syntactic templates in generated text. *arXiv preprint arXiv:2407.00211*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 2024. Toward

self-improvement of llms via imagination, searching, and criticizing. *Advances in Neural Information Processing Systems*, 37:52723–52748.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Peter West and Christopher Potts. 2025. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2024. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *arXiv preprint arXiv:2501.00273*.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020. Trading off diversity and quality in natural language generation. *arXiv preprint arXiv:2004.10450*.

Appendix

A Prompts

This section provides the exact prompts used for data generation, model training, and model evaluation.

A.1 Data Generation Prompts

The prompt used for generating the first response set from the model is as follows,

> **System Prompt:** Task Description: For this task, you will write a very short story. You will be given 3 words, and write a story that includes all 3 words. Your story should be about 5 sentences long. Use your imagination and be creative when writing your story. But, also be sure your story makes sense.
> **User Prompt:** Write a short story that includes these three words: [THREE_WORDS].

The prompt used for generating the second response set from the model is as follows,

> **System Prompt:** Task Description: For this task, you will write a very short story. You will be given 3 words, and write a story that includes all 3 words. Your story should be about 5 sentences long. Use your imagination and be creative when writing your story. But, also be sure your story makes sense.
> **User Prompt:** Write a short story that includes these three words: [THREE_WORDS].
> **Assistant Prompt:** [FIRST_STORY]
> **User Prompt:** I do not like the previous story. Please rewrite the story in the most creative way. The new story: - must be completely different from the previous story in: story plot and characters. - must have a completely different start (do not use standard phrases like "Once upon", "As the", "In a", "In the" etc.). - must be composed of exactly [FIRST_STORY_WORD_COUNT] words. Remember to use the three words: [THREE_WORDS]

A.2 Model Evaluation Prompts

Divergent Association Task The prompt used for the Divergent Association Task (DAT) is as follows,

System Prompt: Task description: Please generate 10 words that are as different from each other as possible, in all meanings and uses of the words. Rules: Only single words in English. Only nouns (e.g., things, objects, concepts). No proper nouns (e.g., no specific people or places). No specialized vocabulary (e.g., no technical terms). Think of the words on your own (e.g., do not just look at objects in your surroundings). Make a list of these 10 words, without any repetition. You must list each word with a number and a period. For example, "1. word-1, 2. word-2, etc."
User Prompt: List 10 words that are as different from each other as possible:

Persona Generation Task (PGT) The prompt used for the Persona Generation Task (PGT) is as follows,

System Prompt: Generate a random persona description with three characteristics. Characteristics are: - First Name - The city of birth - Current occupation Format the output strictly using JSON schema. Use ‘first_name‘ for First Name, ‘city‘ for the city of birth, ‘occupation‘ for current occupation as corresponding JSON keys. The ordering of characteristics should be arbitrary in your answer.

Alternate Uses Task (AUT). The prompt used for the Alternate Uses Task (AUT) is as follows,

System Prompt: Task Description: For this task, you’ll be asked to come up with as many original and creative uses for objects as you can. The goal is to come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different. You must list each use with a number and a period. For example, "1. Use-1, 2. Use-2, 3. Use-3, etc.". You must provide exactly five (5) uses for each object.
User Prompt: Object: [OBJECT], Uses:

The objects used for collecting the AUT responses are as follows,

"belt", "brick", "broom", "bucket", "candle", "clock", "comb", "knife", "lamp", "pencil", "pillow", "purse", "rope", "sock", "table"

Creative Writing Task (CWT). The three-word sets used in evaluating the model are as follows,

("stamp, letter, send"), ("petrol, diesel, pump"), ("statement, stealth, detect"), ("belief, faith, sing"), ("gloom, payment, exist"), ("organ, empire, comply"), ("year, week, embark"),

Instruction for Human Evaluators The instruction provided to the human evaluators is as follows,

```
For each example, you will be given a [PROMPT] and two responses to that prompt: [RESPONSE-1] and [RESPONSE-2]. Your task is to compare the two responses based solely on text diversity—that is, the variety and novelty of vocabulary, sentence structure, and ideas. Do not consider other aspects such as correctness, relevance, or fluency. Choose only one of the following options: A. [RESPONSE-1] is more diverse B. [RESPONSE-2] is more diverse C. Cannot decide Respond with only the letter A, B, or C.
```

Task Description for the LLM-as-a-Judge Evaluation The task description for the LLM-as-a-Judge evaluation is as follows,

```
You are an expert judge of text diversity. For each example, you will be given a [PROMPT] and two responses to that prompt: [RESPONSE-1] and [RESPONSE-2]. Your task is to compare the two responses based solely on text diversity—that is, the variety and novelty of vocabulary, sentence structure, and ideas. Do not consider other aspects such as correctness, relevance, or fluency. Choose only one of the following options: A. [RESPONSE-1] is more diverse B. [RESPONSE-2] is more diverse C. Cannot decide Respond with only the letter A, B, or C.
```

B Pilot Analysis for Sequential Prompting

We conducted an exploratory analysis on $20,000$ short stories generated from Llama-3.1-8B and Olmo-2-7B models (Grattafiori et al., 2024; OLMo et al., 2024). The analysis was targeted at understanding the repeating patterns in the generated stories. With the help of the *diversity* package in Python (Shaib et al., 2024a), we extract the top-5 repeating Part-Of-Speech (POS) bi-grams. We find that the most repeated bigram (*IN DT*) occurs in over 15k stories (out of 20k) and 23% of occurrences are present at the beginning of the generated story, refer to table B.1.

Based on the findings, we conducted a sequential prompting experiment that elicits a more diverse response from the model by asking the model to avoid repeating phrases (refer to appendix A.1 for exact prompts). We find that the diversity of the second response is, on average, higher than the first one.

C Brief Description of ArmoRM score

ArmoRM uses a Llama-3-8B backbone frozen and attaches 19 "meter" heads that score crowd-rated qualities such as creativity, factuality, safety, and verbosity. A small prompt-conditioned gating network assigns non-negative weights to these meters and sums them, after explicitly removing any correlation with length to avoid verbosity bias. In practice the gate down-weights irrelevant meters: on creative-writing prompts $> 70\%$ of the weight falls on creativity-related meters, $< 2\%$ on code-style. The model achieves 89% pairwise accuracy on RewardBench, a benchmark whose preference labels are entirely human-annotated; this already reflects agreement with human judgements.

D Computational Savings Due to Lightweight Filtration

D-NS-Lite was explicitly designed with practical efficiency in mind. Computing entropy and ArmoRM scores for 400,000 generated sequences incurs two additional forward passes per example—one through the base language model and one through the reward model. With an average response length of 150 tokens, each pass processes approximately 60 million tokens. Assuming inference with a 7B-parameter model[9], this results in an estimated 420×10^{15} FLOPs per inference round based on the method outlined in Kaplan et al. (2020), totaling around 840 PFLOPs for both passes. This computational cost—spent solely to measure diversity and quality—represents a significant overhead, roughly equivalent to or exceeding the cost of generating the data itself. As model sizes or dataset scales increase, the cost of metric computation will grow rapidly and can quickly become prohibitive. This motivates the need for low-cost alternatives to expensive metrics like entropy and reward model scores. D-NS-Lite addresses this by replacing these metrics with lightweight

[9]We use Olmo-7B, LLaMA-8B, and ArmoRM, which has approximately 7.5B parameters.

POS Pattern	Example String	Present (out of 20k)	Present at start (%)
IN DT	`As a, In a, In the, At the, On the`	15,782	23.30
DT JJ	`a delicate, the rare, the main, the late`	11,418	16.81
DT NN	`an alley, a monarc, a spoon, a thicket`	18,472	16.03
JJ NN	`single silk, current king, ancient time`	9,335	0.50
NN IN	`hike in, group of, wave of, vendor to`	1,800	0.45

Table B.1: **Repeating bi-grams are more likely at the beginning.** We present the frequency of repeating POS bi-grams. *IN DT* is the most frequent and commonly appears at the start of generated stories.

Metric	First Story	Second Story	Increase in Diversity
TTR	0.7112	0.7469	+0.0357
MAAS (↓)	0.1639	0.1609	+0.0031
HD-D	0.4143	0.4202	+0.0059
MTLD (MA-Bi)	13.9802	14.3997	+0.4195
MTLD (MA)	14.0778	14.5063	+0.4284
MTLD	14.2246	14.6652	+0.4406
MATTR	0.3810	0.3867	+0.0057

Table B.2: **Sequential prompting increases diversity.** We conducted a trial of sequential prompting on $20,000$ responses generated from Llama-8B and Olmo-7B models. The second story generated from the models has higher diversity. ↓: indicates that the lower values of MAAS index represent higher diversity.

lexical proxies such as TTR (Type-Token Ratio) and MAAS. These proxies entirely eliminate the need for the two additional inference rounds, resulting in substantial GPU computation savings. Although proxy metrics are computed on the CPU, their cost is negligible and can be handled efficiently even with modest hardware. Notably, in sequential prompting setups (see Section 4.1), entropy computation for the first set of responses can be avoided, as these are already conditioned on the test-time prompt. However, if one wishes to use later-stage responses to construct the preference dataset, an additional forward pass would be required for each of those responses to match the intended test-time input distribution. By contrast, D-NS-Lite avoids this complication entirely, offering a significantly more efficient pipeline.

E Implementation of DivPO

Brief Description. Lanchantin et al. (2025), similar to our study, propose a data selection strategy for preference tuning to enhance response diversity in base models. Their approach involves two main steps for selecting a preference pair from N responses to a fixed prompt. First, responses are ranked by a text quality score, with the top $\rho\%$ forming the pool of "chosen" candidates and the bottom $\rho\%$ forming the "rejected" pool. Second, each pool is sorted by a diversity metric. A single preference tuning pair is then formed by selecting the most diverse response from the "chosen" pool and the least diverse response from the "rejected" pool.

Hyperparameters. The key hyperparameters in Lanchantin et al. (2025) are the text quality metric, the diversity metric, and the value of ρ. They experiment with rule-based rewards (e.g., JSON validity) and the ArmoRM reward model as quality metrics. In our study, we use ArmoRM exclusively. While they explore various ρ values, we instead define the "chosen" pool as the top quartile (ArmoRM score $\geq 75^{\text{th}}$ percentile) and the "rejected" pool as the bottom quartile (ArmoRM score $\leq 25^{\text{th}}$ percentile). For diversity, they use log probabilities, whereas we use entropy.

F Hyperparameters for Preference Optimization

We fine-tune the base model using the Direct Preference Optimization (DPO) objective (Rafailov et al., 2023), with $\beta = 0.1$ to control the divergence from the original policy. We use a peak learning rate of 1×10^{-5} with a cosine learning rate schedule, and a warm-up phase covering 10% of the total training steps. All models are trained using LoRA adapters (Hu et al., 2021) with a rank $r = 16$ and scaling factor $\alpha = 16$, on a quantized 4-bit backbone model (Dettmers et al., 2023). We add the LoRA modules

to *query* and *value* projection metrics of all transformer layers in the base model with a dropout of 5%.

G Reponse Length Distribution

We observe that the length distribution varies after fine-tuning the model. As presented in table G.1, we observe that the average (and standard deviation) of response length reduces for DivPO and increases for our proposed methods (Diverse-NS and Diverse-NS-Lite). DivPO (inadvertently) teaches the model to generate shorter responses (refer to table 2). Despite maintaining comparable length for "chosen" and "rejected" samples in our methods (Diverse-NS and Diverse-NS-Lite), the model interestingly learns to generate longer responses. We suspect this shift is influenced by a skewed proportion of longer preference pairs, which may inadvertently bias the model toward generating longer responses.

H Results with Standard Deviation

In this section, we report the results with the standard deviation values in Table H.1.

I ΔDD-based Evaluation

Similar to the results presented fig. 1 for Olmo-7B, we present the results for Llama-8B and Olmo-13B in this section.

J A Summary of Metrics

We provide a concise summary of all metrics used in our evaluation setup in Table J.1.

K LLM-as-a-Judge Evaluation on CWT

We extend our human evaluation setup to assess diversity gains using the LLM-as-a-Judge framework. Similar to the human study, we present a single prompt and two model responses to an LLM and ask it to select the more diverse output (see Appendix A.2 for prompts). For each CWT prompt, we generate 10 responses per method, yielding 100 unique pairs for direct comparison. To ensure maximally distinguishable outputs, we select the 50 pairs with the lowest Jaccard similarity. Using 7 prompts, we construct 350 evaluation pairs per method comparison. Each comparison is evaluated by five proprietary LLMs: Claude-3.5-Haiku, Claude-Sonnet-4, GPT-4.1, GPT-4.1-mini, and GPT-4o-mini. We report the win percentages for D-NS and D-NS-Lite against the base model, DivPO, and each other in Table K.1. The results align with human evaluation findings: D-NS outperforms both baselines in all but one case, and D-NS-Lite achieves similar improvements, also outperforming the baselines in all but one setting. The comparison between D-NS and D-NS-Lite is more competitive, with win rates hovering around 50%, slightly favoring D-NS.

L Examples of Model Responses for CWT Prompts

In this section, we provide a couple of examples of the model responses to CWT prompts, before and after the diversity tuning. Please refer to the Figure L.1.

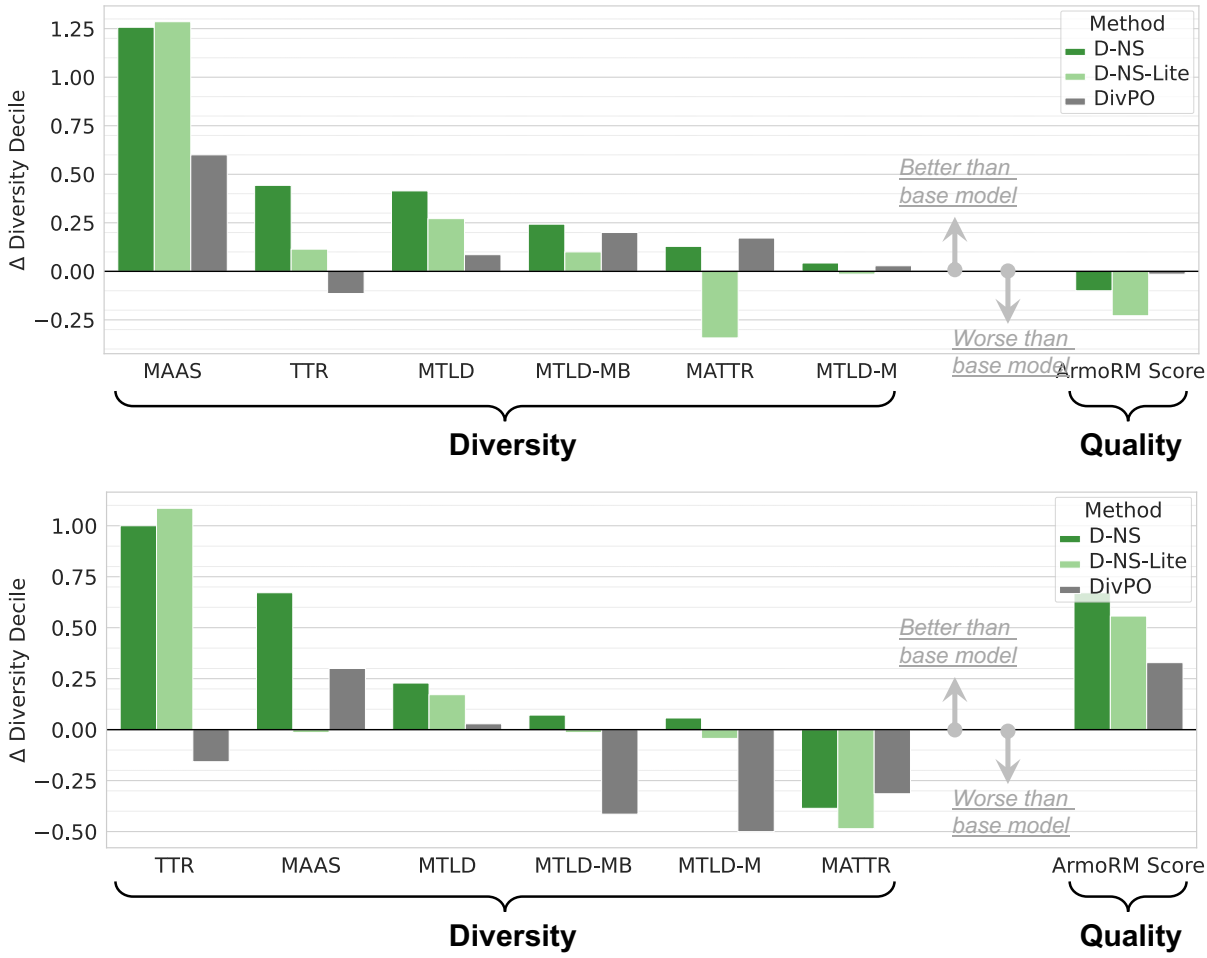

Figure I.1: **Diversity and Quality Evaluation on CWT.** This figure shows ΔDiversity Decile (ΔDD) values (y-axis) across various metrics (x-axis), computed from 70 CWT responses generated by the Llama-8B model (top-panel) and Olmo-13B (bottom panel). A value of zero represents base model performance; bars indicate improvements from preference-tuned models.

Model	Base Model	DivPO (Lanchantin et al., 2025)	Ours - D-NS-Lite	Ours - D-NS
Llama-8B	123.27 ± 18.14	111.24 ± 14.89	141.44 ± 37.26	139.47 ± 33.65
Olmo-7B	73.63 ± 15.47	62.27 ± 12.88	81.37 ± 18.21	83.91 ± 17.93
Olmo-13B	86.11 ± 13.96	72.20 ± 13.87	101.40 ± 17.64	100.60 ± 18.01

Table G.1: **Change in the Response Length.** In this table, we present the average length of model-generated responses before and after the preference-tuning. The average values are calculated on 70 responses generated on the CWT evaluation prompts.

Task	Metric	Base Model	DivPO	D-NS-Lite	D-NS
		LLaMA-8B			
DAT	DSI	0.7535 ± 0.07	0.7545 ± 0.06	0.7590 ± 0.07	**0.7640** ± 0.07
DAT	Unique Words	0.4575	0.4593	0.4797	**0.4914**
PGT	Unique First Names	0.6500	0.6100	**0.6900**	**0.6900**
PGT	Unique Cities	0.3300	0.3100	**0.4700**	0.4200
PGT	Unique Occupations	0.4100	0.3900	**0.5100**	**0.4900**
AUT	DSI	0.8876 ± 0.02	0.8837 ± 0.02	0.8876 ± 0.02	**0.8878** ± 0.02
CWT	DSI	0.8515 ± 0.01	0.8521 ± 0.01	0.8556 ± 0.01	**0.8581** ± 0.01
CWT	ArmoRM Score	0.1451 ± 0.02	**0.1495** ± 0.01	0.1369 ± 0.02	0.1405 ± 0.02
CWT	4-gram div. POS	0.4990	0.4990	**0.5030**	0.5000
CWT	4-gram div.	2.8550	2.9320	2.9450	**2.9620**
CWT	Comp. Ratio.	2.635	2.546	2.568	**2.530**
		OLMo-7B			
DAT	DSI	0.7480 ± 0.09	0.7509 ± 0.08	**0.7662** ± 0.08	0.7639 ± 0.08
DAT	Unique Words	0.6139	0.6079	**0.6347**	0.6327
PGT	Unique First Names	0.3300	0.3300	0.3300	**0.3400**
PGT	Unique Cities	**0.3100**	0.3000	0.2700	0.2700
PGT	Unique Occupations	0.5200	0.5500	**0.6100**	**0.6100**
AUT	DSI	0.8836 ± 0.02	0.8846 ± 0.02	0.8852 ± 0.02	**0.8858** ± 0.02
CWT	DSI	0.8499 ± 0.01	0.8491 ± 0.01	0.8548 ± 0.01	**0.8563** ± 0.01
CWT	ArmoRM Score	0.1435 ± 0.02	0.1441 ± 0.02	0.1462 ± 0.01	**0.1464** ± 0.01
CWT	4-gram div. POS	0.5720	**0.5770**	0.5350	0.5530
CWT	4-gram div.	3.1270	3.1690	**3.1750**	3.1620
CWT	Comp. Ratio.	2.4460	2.4160	**2.3850**	2.3970
		OLMo-13B			
DAT	DSI	0.7233 ± 0.06	0.7282 ± 0.07	0.7320 ± 0.06	**0.7364** ± 0.06
DAT	Unique Words	**0.3421**	0.3340	0.3310	0.3256
PGT	Unique First Names	0.4100	0.4100	0.4400	**0.4500**
PGT	Unique Cities	0.3500	0.3500	0.3700	**0.3900**
PGT	Unique Occupations	0.1900	0.1900	0.1900	**0.2000**
AUT	DSI	0.8943 ± 0.02	0.8960 ± 0.02	**0.8974** ± 0.02	0.8970 ± 0.02
CWT	DSI	0.8557 ± 0.01	0.8555 ± 0.01	**0.8616** ± 0.01	0.8614 ± 0.01
CWT	ArmoRM Score	0.1571 ± 0.01	0.1589 ± 0.01	0.1585 ± 0.01	**0.1590** ± 0.01
CWT	4-gram div. POS	0.5210	**0.5229**	0.5080	0.4960
CWT	4-gram div.	3.0820	3.0770	3.095	**3.1070**
CWT	Comp. Ratio.	2.492	2.512	2.505	**2.480**

Table H.1: **Diversity and Quality Evaluation.** We present the average (± std. dev.) diversity (DSI or unique values) and quality (ArmoRM score) measurements for model responses collected on four creative generation tasks (Structured Gen.: DAT, PGT, Free-Form Gen.: AUT, CWT).

Metric	Definition	Trend Description (Trend)	Application
Entropy	Entropy of the token distribution in a response; measures unpredictability.	Higher values indicate greater lexical diversity (↑).	Training-data filtering and diversity-bias analysis
Type–Token Ratio (TTR)	Ratio of unique token types to total tokens.	Higher values indicate more lexical variety (↑).	Lightweight filtering (D-NS-Lite), Calculation of Diversity Decile
Moving-Average TTR (MATTR)	Moving-average of TTR over sliding windows; smooths variability.	Higher values indicate greater lexical diversity (↑).	Correlation analysis, Calculation of Diversity Decile
Measure of Textual Lexical Diversity (MTLD)	Average segment length until TTR falls below a threshold; longer segments imply more diversity.	Higher values indicate greater lexical diversity (↑).	Correlation analysis, Calculation of Diversity Decile
Moving-Average MTLD (MTLD-M)	Moving-average smoothing of MTLD to reduce variance.	Higher values indicate greater lexical diversity (↑).	Correlation analysis, Calculation of Diversity Decile
Bidirectional Moving-Average MTLD (MTLD-MB)	MTLD-M applied forward and backward for context-sensitive smoothing.	Higher values indicate greater lexical diversity (↑).	Correlation analysis, Calculation of Diversity Decile
MAAS	Proxy metric correlated with ArmoRM quality scores.	Higher values indicate stronger quality/diversity signal (↑).	Lightweight filtering (D-NS-Lite), Calculation of Diversity Decile
Hypergeometric Distribution Diversity (HD-D)	Probability-based measure of lexical diversity under a hypergeometric model.	Higher values indicate greater lexical diversity (↑).	Correlation analysis
ArmoRM score	Holistic quality score from a reward model.	Higher values indicate better fluency–diversity trade-off (↑).	Quality evaluation (Creative Writing Task) and filtering, Calculation of Diversity Decile
Divergent Semantic Integration (DSI)	Average semantic distance among items in a generated list.	Higher values indicate greater divergent thinking (↑).	Diversity evaluation (Divergent Association Task, Creative Writing Task)
Diversity Decile (DD)	Decile rank of a response's diversity within its length group.	Higher decile indicates higher relative diversity after length normalization (↑).	Length-normalized evaluation (Creative Writing Task)
Change in Diversity Decile (ΔDD)	Difference in DD before and after tuning; quantifies diversity gain.	Positive values indicate diversity gain; negative indicate loss (↑/↓).	Measuring tuning effect on diversity (Creative Writing Task)
Semantic Distance (SD)	Average embedding-space distance between outputs; indicates semantic variety.	Higher values indicate greater semantic variety (↑).	Diversity evaluation (Alternate Uses Task)

Table J.1: **Overview of diversity and quality metrics:** definitions, trend descriptions with arrows, and their applications including evaluation tasks.

Judge LLM	**D-NS vs. Base** D-NS Win %	**D-NS vs. DivPO** D-NS Win %	**D-NS vs. D-NS-Lite** D-NS Win %
claude-3-5-haiku	54.86	66.86	47.43
claude-sonnet-4	52.86	49.43	54.57
gpt-4.1	74.86	82.86	54.86
gpt-4.1-mini	61.43	78.86	48.57
gpt-4o-mini	75.14	82.86	51.43
Judge LLM	**D-NS-Lite vs. Base** D-NS-Lite Win %	**D-NS-Lite vs. DivPO** D-NS-Lite Win %	**D-NS-Lite vs. D-NS** D-NS-Lite Win %
claude-3-5-haiku	58.57	63.71	52.00
claude-sonnet-4	48.86	51.14	45.14
gpt-4.1	77.43	81.71	45.14
gpt-4.1-mini	68.57	78.00	37.14
gpt-4o-mini	76.57	81.43	48.57

Table K.1: **LLM-as-a-Judge Evaluation on CWT.** Pairwise LLM-as-a-Judge evaluation results comparing D-NS and D-NS-Lite against the base model, DivPO, and each other on the CWT task responses. Each cell presents the win percentage values. D-NS and D-NS-Lite achieve higher win rates against both baselines in all but two cases. The win rate values for D-NS versus D-NS-Lite comparison hover around 50% while being slightly in favor of D-NS.

Prompt: Write a creative story that includes the following words: belief, faith, sing.
Base model response: Once upon a time in a small town, belief and faith walked arm and arm. A man believed he could sing, despite having never uttered a note. One day, he decided to test the faith he had in his voice by entering a talent show. He sang with his heart, and to his amazement, the whole town believed his voice carried melody. His belief, supported by faith, made him that much more willing to sing, earning him first place and respect from all.
Response after D-NS tuning: In a quaint little village by a lush river, lived a community bound by belief and faith that a melody held magical powers. Every full moon, they gathered to sing under the shimmering silver light, believing their voices would protect them and bring prosperity to their lands. Faith, the backbone of their society, was unwavering despite the growing skepticism outside their village. That night, as they sang together, a gentle wave turned into a miraculous surge of bounty, renewing their faith and proving their belief in the power of song. The mesmerizing melody danced with the winds, a testament to their harmonious bond and shared beliefs.

Prompt: Write a creative story that includes the following words: year, week, embark.
Response after DivPO tuning: A new year brought new hope to the small town. Every resident embarked on a journey to make the year their best yet. One week in, Lucy decided to lose weight and improve her health. Her whole year proceeded as she planned, full of determination and hard work. By the end of the year, she had lost fifty pounds and was radiant with health.
Response after D-NS tuning: Every year, during the week of Christmas, a group of friends would embark on a magical adventure through the snowy town. Each year, their journey was marked by a whimsical event that seemed to defy reality, like talking animals or a melody that could heal the sick. This year, as winter approached, their anticipation was high. The week finally came, and, true to form, they discovered an ice-encased river that sang an ancient melody under the moonlight. Together, they ventured onto the ice, music echoing around them, and the frosty air filled their lungs, a reminder of magic's fleeting touch in our everyday year.

Figure L.1: **Model Responses Before and After Diversity Tuning.**