From Macro to Micro: Probing Dataset Diversity in Language Model Fine-Tuning

Haoyu Li[1*], **Xuhong Li**[2*], **Yiming Dong**[3], **Kun Liu**[1†]

[1]School of Automation, Beijing Institute of Technology
[2]Baidu Inc.
[3]School of Physics, Peking University
haoyli_bit@bit.edu.cn, lixuhong@baidu.com, ydong@pku.edu.cn, kunliubit@bit.edu.cn

Abstract

Dataset diversity plays a pivotal role for the successful training of many machine learning models, particularly in the supervised fine-tuning (SFT) stage of large language model (LLM) development. Despite increasing recognition of its importance, systematic analyses of dataset diversity still remain underexplored. To address this gap, this work presents a systematic taxonomy of existing diversity-control strategies, which primarily focus on the *instruction* component, operating at either macroscopic (entire instruction semantics) or mesoscopic levels (instruction units), and furthermore introduces a novel analysis of microscopic diversity within the *response* component, specifically analyzing the statistical distribution of **tokens** in SFT training samples. In the experimental evaluation, we construct fixed-size datasets (e.g., 10,000 samples each) from a corpus of 117,000 open-source SFT samples, incorporating six distinct diversity-control strategies spanning macro-, meso-, and microscopic levels applied to both instructions and responses. We then fine-tune LLMs on these datasets to assess the six diversity-control strategies. Results reveal that while macroscopic and mesoscopic strategies lead to higher performance with increasing diversity, the microscopic strategy in responses exhibits not only a stronger correlation between model performance and the degree of diversity, but also superior performance with maximum diversity across all strategies. These findings offer actionable insights for constructing high-performance SFT datasets.

Code — https://github.com/li-haoyu/M2M-Diversity
Extended version — https://arxiv.org/abs/2505.24768

1 Introduction

The success of large language models (LLMs) hinges on the advancements in model architectures and computational resources, and more critically on the acquisition and management of training data (Kaplan et al. 2020; Ouyang et al. 2022; Achiam et al. 2023; Wang et al. 2023). While significant research has focused on the quality of training samples, increasing attention is being given to dataset diversity that enhances models' generalizability to handle real-world scenarios (Bubeck et al. 2023; Bukharin et al. 2024; Zhao et al. 2024a; Zhou et al. 2023). This work advances this line of research and focuses on the dataset diversity during the supervised fine-tuning (SFT) stage of the LLM alignment.

Existing diversity-control strategies prove to be effective in enhancing LLM capacity. The first strategy is to cluster data samples by the semantics of instructions (Grootendorst 2022; Du, Zong, and Zhang 2023) or instruction-response pairs (Ge et al. 2024b), and improve the dataset diversity by utilizing more clusters as possible. An alternative strategy, exemplified by InsTag (Lu et al. 2023), further decomposes the instruction into atomic components, and increases the dataset diversity by covering more instruction unit tags.

While these strategies prove to be effective, there are two challenges along this research direction. First, systematic analyses and metrics of dataset diversity still remain underexplored. Different datasets, models and evaluations are used in previous works, and few metrics can correlate dataset diversity to the model performance significantly. Secondly, previous works mainly focus on the *instruction* component of the instruction-response pair in the SFT dataset for diversity control. Note that LLMs are conventionally supervised using *responses* as the primary training signals (Brown et al. 2020; Hu et al. 2022; Rafailov et al. 2023). Though instructions serve as indicators of the diversity of topics, domains, disciplines or other semantic aspects (Wei et al. 2022; Wang et al. 2023; Zhou et al. 2023) and may implicitly diversify responses, explicit signals might be more effective.

To tackle the above challenges, we first present a taxonomy on the diversity strategies from macro- (entire instructions), meso- (instruction units) and microscopic (tokens) levels and on both instruction and response components during the SFT stage of LLM training. We then propose to explicitly examine the impact of controlling diversity within the *response* component of SFT datasets and compare its effectiveness to that of the instruction component. For comprehensive comparisons, we apply various diversity control strategies to construct SFT datasets from 117K open-source instructions where the responses are re-constructed for quality control, and train hundreds of models using Llama series models (Touvron et al. 2023; Dubey et al. 2024) for quantifying the effectiveness of different diversity strategies.

Despite these efforts, defining dataset diversity and establishing robust metrics correlating diversity with model performance remain challenging. We collect multiple diversity-

[*]These authors contributed equally.
[†]Corresponding author.

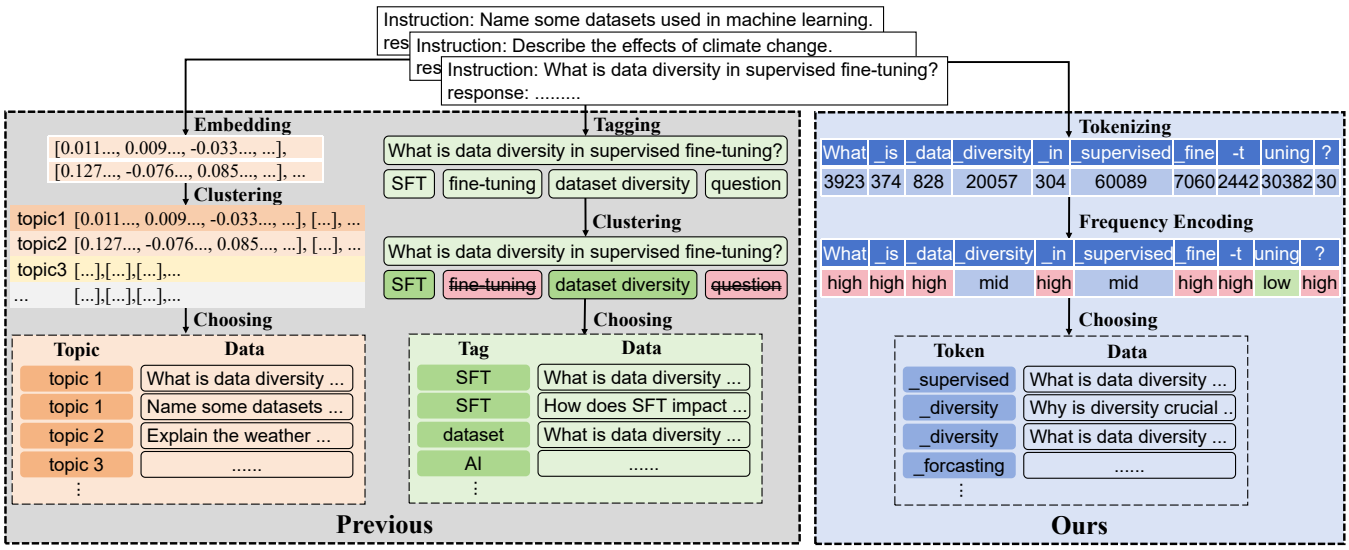


Figure 1: Diversity-control strategies across three scales on instruction. At the macroscopic scale, each instruction is assigned to a corresponding topic after embedding and clustering (Grootendorst 2022). At the mesoscopic scale, each instruction is linked to multiple relevant tags through LLM-based tagging, filtering, and clustering (Lu et al. 2023). At the microscopic scale, we tokenize all instructions and select representative tokens based on their frequencies across the corpus for each instruction. Diversity-controlled datasets are then constructed based on topic/tag/token varieties, which also serve as diversity indicators. The same diversity-control strategies are also applied to the **response** perspective.

related metrics and involve an additional metric relating to information theory, which may serve as a powerful tool for quantifying dataset characteristics. We present analyses and discuss the advantages and limitations of these metrics. Specifically, the contributions of this work can be summarized as follows:

(1) A systematic taxonomy on dataset diversity-control strategies is proposed, from macro-, meso- to microscopic levels on instruction and response components of SFT datasets.

(2) A novel proposed approach on controlling the response diversity from the token (microscopic) level exhibits both a stronger correlation between model performance and the degree of diversity and superior performance with maximum diversity across all strategies.

(3) Comprehensive experiments are conducted involving fix-size datasets, incorporating six distinct diversity-control strategies (spanning three levels, applied to both instructions and responses). Detailed analyses across three dataset sizes, three strategies, two components, multiple diversity metrics, are also provided. All results demonstrate that macroscopic and mesoscopic strategies effectively enhance model performance with increasing diversity, while the microscopic strategy on responses not only displays stronger performance-diversity correlation, but also achieves the optimal model performance among all strategies when diversity is maximized.

2 Methodology and Framework

In this section, we introduce the three-level perspective of dataset diversification strategies. We first review the macro- and mesoscopic ones on the instruction component in Section 2.1. Then we present the proposed strategy of the microscopic one on the response component in Section 2.2. The illustration of previous works and our proposed method is shown in Figure 1. We eventually introduce the framework of empirical comparisons and analyses across all strategy combinations, including data source, dataset construction, model training evaluations and the diversity metrics, in Section 2.3.

2.1 Previous Works: Macro- and Mesoscopic Diversity Strategies

Macroscopic analysis of dataset diversity focuses on the semantic content of texts to characterize thematic diversity. BERTopic (Grootendorst 2022) is one typical approach of this strategy and employs the following steps: (1) Embedding the texts into dense vector representations via any tokenizer (e.g., (Mikolov et al. 2013)); (2) Clustering embeddings to semantically related clusters (e.g., HDBSCAN (McInnes, Healy, and Astels 2017)), where dimension-reduction techniques may be helpful (e.g., UMAP (McInnes et al. 2018)). Then the dataset diversity scale can be controlled by choosing samples from a certain amount of clusters.

Multiple approaches (Du, Zong, and Zhang 2023; Ge et al. 2024b,a) can be categorized into this macroscopic strategy with various differences. For example, CaR (Ge et al. 2024b) additionally ranks the samples within each cluster and selects the top-quality samples. MoDS (Du, Zong, and Zhang 2023) computes the embeddings and selects examples by maximizing the embedding distance. A persona-based method (Ge et al. 2024a) has been proposed to create synthetic instructions and maximize the diversity of instructions.

Mesoscopic analysis of dataset diversity focuses on decomposing an entire text into several unit tags, and then

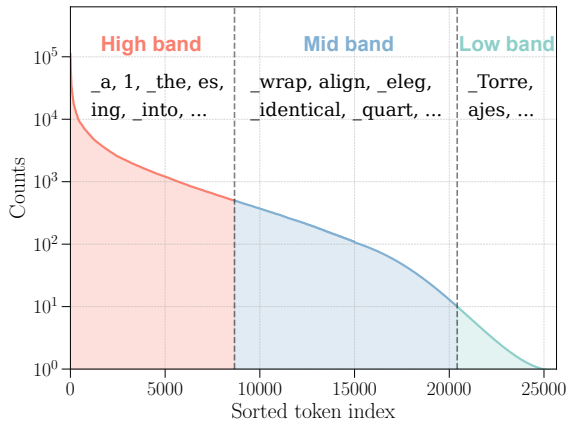


Figure 2: Distribution for tokens in the SFT dataset. Based on the counts of tokens, we classify them into three categories: "High band" (more than 500 counts), "Mid band" (10-500 counts), and "Low band" (fewer than 10 counts), and present several examples for each scenario.

clustering the tags instead of the entire semantic content of texts. InsTag (Lu et al. 2023) is the typical approach, which is achieved by the two-step processing approach: (1) Generating unit tags for each text via an LLM; (2) Clustering tags into semantically coherent groups. Similar to macroscopic strategy, the dataset diversity can be also controlled by including different amounts of tags.

2.2 Our Proposed Method: The Microscopic Strategy on Responses

While macro- and mesoscopic analyses provide valuable insights into the dataset diversity, existing approaches predominantly concentrate on *the instruction component*. Although instructions act as proxies for contextual diversity, they are not an explicit signal as the LLMs are supervised on responses only in convention during the SFT stage. This fundamental misalignment motivates our dual innovation: (1) a paradigm shift from instruction-centric to response-driven diversity analysis, and (2) the introduction of a microscopic characterization framework operating at the token granularity. Specifically, the microscopic strategy on *the response component* involves the following two steps:

(1) Tokenizing Texts. The proposed method begins by tokenizing all responses (or instructions, without loss of generality) using an LLM's tokenizer. Token frequency, defined as *the number of occurrences of a token that appears in the training samples*, is then calculated. A manual inspection of tokens across the frequency spectrum leads to their classification into three bands: high, mid, and low (Figure 2). High-band tokens are typically prepositions, articles, letters, numbers, and common prefixes and suffixes, which often lack specific semantic meaning. Conversely, low-band tokens tend to be names, loanwords, or typos. Consequently, *mid-band tokens* carry most of the semantic meaning for the texts, and are flagged for subsequent processing.

(2) Controlling the Dataset Diversity at Token-Level. Controlling the diversity at the token level is more challenging than the query or tag levels. One sample may contain over 20 mid-band tokens on average, compared to one semantic cluster or four tags. This quantitative complexity leads to a non-uniform dataset distribution, causing a problem that cannot be trivially resolved by sampling samples from a predefined number of clusters or tags, as previous works do, because the number of token types cannot be controlled in such way. We therefore propose a two-stage algorithm to control the number of clusters and samples separately. Specifically, Algorithm 1 prunes the dataset by greedily removing samples with the most unique token types until a target number of token types is reached. Algorithm 2 then selects a preset number of samples, controlling the data distribution while ensuring no token types are lost. The pseudocode for Algorithm 1 and 2 is provided in Appendix D.

2.3 Comprehensive Framework of Comparison

To systematically compare previous diversity strategies and our proposed approach, we propose a taxonomy and a framework to contain a coherent dataset construction, model training and evaluation pipeline with the dataset diversity strategy modifiable. Within the proposed framework, we conduct experiments to validate the effectiveness of our approach and advance this line of research.

Training Data Preparation We compile a comprehensive dataset from a variety of open-source resources. After a rigorous cleaning process, which includes deduplication and filtering based on character encoding, a refined instructional set of 117K prompts is left. To alleviate the quality variance, we ignore the original responses and reconstruct the paired responses by prompting Llama-3.1-70B-Nemotron model[1] (Wang et al. 2024), to ensure that training samples are at a similar level of quality, without forgetting that the decoding strategy still involves some variance. Dataset resources used in this work are documented in Appendix A.

From the same data source, we employ a diversity strategy to construct a series of SFT datasets spanning from minimal to maximal diversity under this strategy. We maintain fixed dataset sizes across experiments while exploring three distinct scales: 10K, 20K, and 30K samples per experimental set. All dataset construction strategies in this work culminate in a uniform selection process, resulting in datasets formally defined as: $\mathcal{D}_k = \bigcup_{i=1}^{k} \Pi(\mathcal{C}_i)$, where k is the number of topic/-tag/token types controlling the diversity scale, $\mathcal{C}_i$ refers to the set of samples regarding to the i-th type, and Π represents the sampling strategy designed to achieve a uniform distribution across target types in $\mathcal{D}_k$. This generates a series of datasets for each diversity strategy denoted as $\{\mathcal{D}_m\}_{m=k_1,k_2,\cdots,k_M}$, where k_1 and k_M specify the minimal and maximal diversity bounds, M indicates the number of datasets in the series (typically set to 7 in our experiments), and intermediate k_i values can be algorithmically determined.

Model Training and Evaluations Each series of datasets is then fine-tuned from the same pretrained model, where

[1]It is the best open-source model on lmarena at the time when starting this work.

Metric	Equation	Explanation
N-gram Ratio (NR) ↑	$R_n = \frac{\#\text{Unique } n\text{-grams}}{\#\text{Total } n\text{-grams}}$	The ratio of unique n-grams to total n-grams, serving as a measure of lexical diversity (Padmakumar and He 2024).
Embedding Distance (ED) ↑	$D_{\text{avg}} = \frac{1}{N}\sum_{a\neq b}\|\mathbf{v}_a - \mathbf{v}_b\|$	The average distance between embeddings (Arora, Liang, and Ma 2017), where $\mathbf{v}$ is the embedding of an element.
Sequence Length (SL) ↑	$L_{\text{seq}} = \frac{1}{N}\sum_{i=1}^{N} l_i$	The mean number of tokens over sequences (Zhao et al. 2024b).
Compression Ratio (CR) ↑	$C_{\text{ratio}} = \frac{\text{Original size}}{\text{Compressed size}}$	The ratio of original dataset size to compressed size (Shaib et al. 2024).
Self-BLEU (BL) ↓	$S_{\text{BLEU}} = \frac{1}{N}\sum_{i=1}^{N}\text{BLEU}_{-i}$	The average $BLEU_{-i}$ score where each text i is compared against the rest of the dataset $(-i)$ (Zhu et al. 2018).
Information Entropy (IE) ↑	$E_{\text{Entro}} = -\sum_{i=1}^{n} p_i \log(p_i)$	The measure of the randomness of the distribution, where p_i is the probability/frequency of a token.

Table 1: Posterior diversity metrics. The upward-pointing arrow denotes a positive correlation between the metric value and dataset diversity, while the downward-pointing arrow indicates an inverse relationship (negative correlation).

Llama-2-7B (Touvron et al. 2023) is used for most of our fine-tuning experiments. For the evaluations, we adopt the pairwise scoring methodology used in Arena Hard (Li et al. 2025) and combine the testing samples of both AlpacaEval 2.0 and Arena Hard benchmarks (Li et al. 2023) for a more comprehensive evaluation. GPT-4 Turbo is originally used by the Arena Hard scoring system. However, due to its high cost, we replace the judge with Llama-3.1-70B-Nemotron (Wang et al. 2024), which shows a very high degree of agreement, with a reversal rate of only 8%. Additional evaluations using alternative judge models further confirm this observation. Further details regarding the evaluation setup and the consistency analysis are provided in Appendix B.

At the end of this step, we obtain the scores $\mathcal{S}$ with respect to constructed datasets, *i.e.*, $\{\mathcal{D}_m, \mathcal{S}_m\}_{m=k_1,k_2,\cdots,k_M}$. Results on these two items across multiple diversity scales (M), three dataset sizes (10K, 20K and 30K), three levels(macro-, meso- and microscopic) and two components (instructions and responses), as well as detailed analyses and ablation studies, will be introduced in the following two Sections.

Posterior Diversity Metrics Previous works have proposed to measure the diversity by the metrics listed in Table 1 that provide quantitative insights into different aspects in posterior, while some of them are used or can be optimized when constructing datasets. Moreover, we would like to mention the metric of the information entropy that may be a reasonable metric. Detailed discussions are provided in Section 4.4.

3 Main Results

In this section, we conduct the comparative experiments as introduced in the previous section, to compare the tuples of datasets and the corresponding scores. To briefly recall the framework, there are three aspects that experiments will be conducted for the comparison:

(1) Across multiple diversity scales. For a certain size of datasets (e.g., 10K), we vary the diversity scale from the lowest to the highest that the strategy can achieve. This range of diversity defines the bounds, where we manually set the lowest as 0% and the highest as 100%. The values in-between represent different diversity scales. This definition makes it possible to compare across strategies.

(2) Across three dataset sizes. We choose three sizes of constructed datasets, *i.e.*, 10K, 20K and 30K from a corpus of 117K instructions and refreshed responses.

(3) Across three strategic levels and two components. We apply macro-, meso- and microscopic dataset diversity strategies on instructions and responses, to perform a comprehensive comparison.

At the end of this section, we furthermore test the effectiveness of the listed posterior diversity metrics for measuring the correlation between diversity and performance.

3.1 Comparative Results

We show the main results in Figure 3, where the observations and findings are discussed below. Note that as the pairwise scoring system requires a reference baseline, a random 10K dataset is used for all experiments in this section. An equal model scores 50 according to this scoring system, so dashed gray lines are drawn at the score of 50.

Dataset Sizes Our experiments with constrained dataset sizes (10K, 20K and 30K training samples) reveal a consistent performance trend where models trained on larger datasets generally outperform those using smaller training datasets. While we hypothesize that diminishing returns or performance plateaus might emerge at greater dataset magnitudes, systematic investigation of this phenomenon remains beyond the scope of the current study.

Nevertheless, increased dataset diversity helps mitigate the performance gap associated with varying dataset sizes. Notably, some smaller yet more diverse datasets curated on the response component at the token level achieve superior model performance compared to larger yet less diverse datasets. This finding underscores the importance and effectiveness of dataset diversity during dataset construction.

Macro-to-Micro Diversity Strategies The statistical results in Table 2, which quantify the slopes of lines between the semantic diversities and scores of fine-tuned LLM models

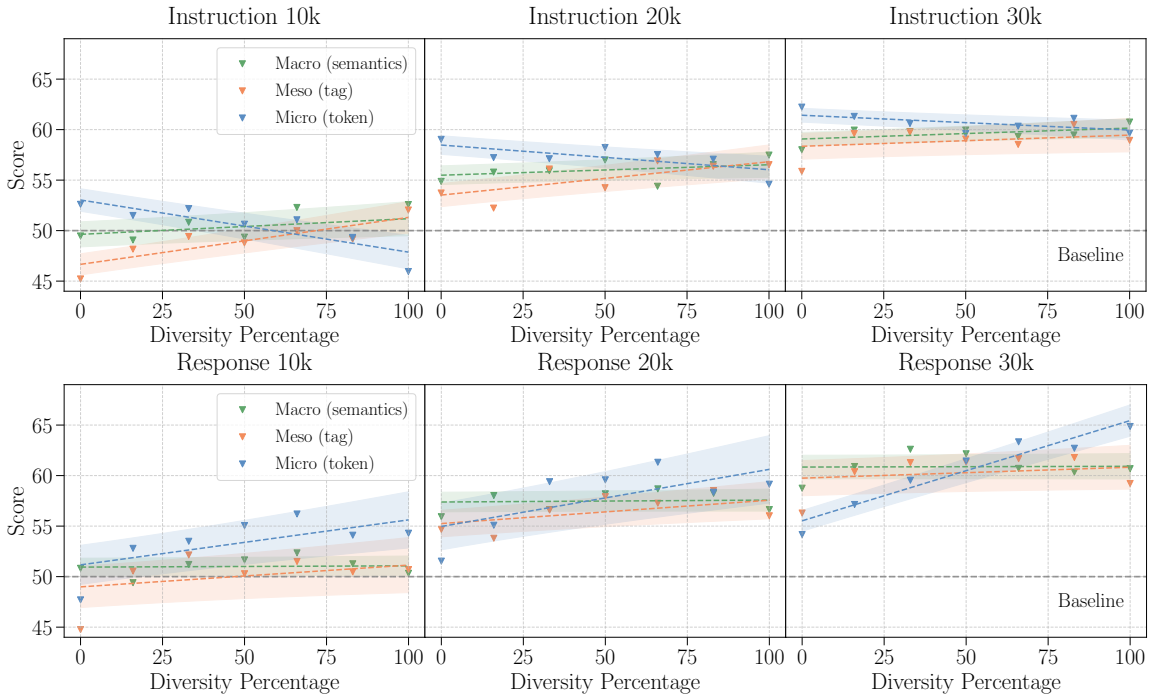


Figure 3: The relationship between the diversity percentage and model performance for instructions and responses across three dataset sizes (10K, 20K, and 30K). The vertical axis shows the average score from two benchmarks, with a baseline score of 50, derived based on a randomly selected 10K dataset. In each subplot, different colors correspond to three diversity levels: Macro (semantics), Meso (tag), and Micro (token). The markers represent the data, while the dashed lines and shaded areas indicate the results of Bayesian linear regression with 1-σ uncertainty.

Dataset	Instruction			Response		
Size	Macro	Meso	Micro	Macro	Meso	Micro
10K	2.37	4.93	−5.45	0.73	3.61	**5.35**
20K	1.67	3.75	−2.83	0.97	1.71	**6.68**
30K	1.45	2.07	−1.80	0.59	2.00	**10.06**

Table 2: Slope ($\times 10^{-2}$) of the linear regression fit in Figure. 3 for each diversity-control strategy.

(as depicted in Figure 3), align with the hypothesis that instruction sets encompassing broader semantic ranges enhance a model's generalization capability. Specifically, we observe that the **macroscopic** strategy (semantics level) on instructions show positive correlations between ranges enhance a model's generalization capability. A similar yet weaker correlation is observed for the response component. However, compared to another two strategies, the macroscopic strategy achieves a smaller range of performance scores, suggesting it is less effective than finer-grained strategies.

Results of the **mesoscopic** strategy (tag level) show consistently positive correlation between the diversity scales and scores, on both instructions and responses. The mesoscopic strategy shows larger slopes than the macroscopic, demonstrating that decomposing the entire instruction into functional attributes of the text, such as topics, intents, sentiments, areas *etc* is more effective.

The proposed **microscopic** strategy (token-level) demonstrates limited efficacy when applied to instructions but exhibits the most significant impact on responses compared to the previous two strategies. This differential performance aligns with expectations. The effectiveness on responses stems from the fact that LLMs are explicitly supervised on response tokens during the SFT stage. Since LLMs generate outputs autoregressively, introducing token diversity likely mitigates overfitting by forcing the model to generalize across varied tokens, making training more robust and generalizable.

Conversely, the strategy's failure on instructions is intuitively explainable. First, token-level diversity in instructions does not inherently translate to meaningful semantic diversity. A microscopic focus on tokens risks overlooking broader contextual and semantic relationships. Second, we hypothesize that successful LLM alignment depends on consistent exposure to recurring token patterns in instructions. Reducing the frequency of specific tokens (via diversification) may dilute critical alignment signals, weakening the model's ability to internalize task-specific linguistic or behavioral norms.

3.2 Tests of Diversity Metrics

Most of the investigated metrics, as detailed in Section 2.3 and Table 1, cannot be easily optimized during the dataset construction, and thus require to be tested in a posterior way. For experiments per dataset size, we compute the metric

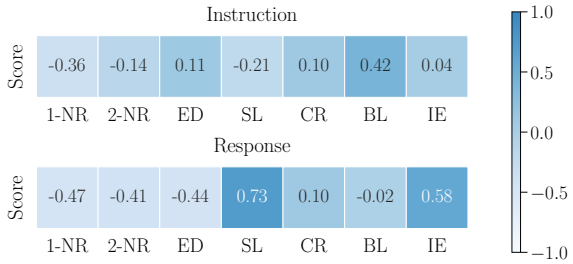


Figure 4: Pearson correlation coefficients of multiple diversity parameters and model performance scores based on all the involved 10K datasets. The upper and lower columns represent the diversity measurements from the perspectives of instruction and response, respectively. The intersection of two parameters shows their Pearson correlation coefficients.

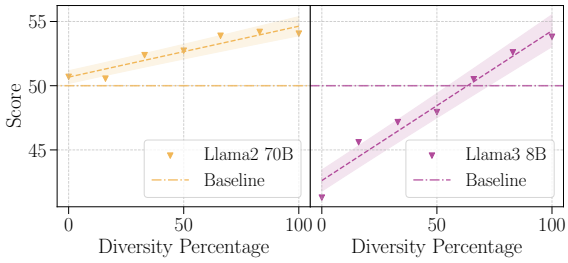


Figure 5: Ablation study on microscopic diversity: comparing different models (Llama-2-70B and Llama-3-8B). The abscissa indicates the microscopic level diversity percentage from the response perspective in the case of 10K dataset.

values and measure the correlation to performance scores. Figure 4 shows the results. Note that the metrics are applied to the instruction and the response respectively.

For the instruction, the most correlated metric is Self-BLEU (0.42). This coheres with the macroscopic strategy on instructions. Both would like to reserve diverse instructions, while the macroscopic applies the semantically clustering way and Self-BLEU measures the semantic similarities. However, as noticed previously, the macroscopic strategy is less effective than meso- and microscopic strategies.

Turning to response-level metrics, sequence length (0.73) and information entropy (0.58) exhibit the strongest correlations with performance scores. The high correlation of sequence length aligns with prior findings that prioritizing longer responses serves as an effective heuristic for enhancing model performance in SFT datasets (Zhao et al. 2024b). To investigate whether this correlation reflects direct causation, we conduct ablation experiments in Section 4.2 using the proposed microscopic strategy on responses filtering with strict length constraints. By maintaining near-identical lengths across compared datasets, the ablation study isolates length as a variable to assess its impact on model outcomes.

The information entropy metric relates to the core mechanism of our microscopic strategy, with the slight difference that the strategy optimizes the entropy over the set of mid-band tokens and the metric computes over the whole vocabulary. The strong correlation with information entropy implies the effectiveness of the microscopic strategy. Since entropy measures the distribution of training signals when computed on responses, high entropy indicates diverse tokens. This helps reduce overfitting and promotes more robust, generalizable training. Detailed analyses are in Appendix E.

4 Ablation Studies

Main experimental results indicate that macroscopic and mesoscopic strategies effectively lead to higher performance with increasing diversity, whereas the microscopic strategy on response exhibits both a stronger correlation between model performance and diversity, and superior performance with maximum diversity across all strategies. To further investigate the mechanisms of the proposed strategy, we conduct three ablation studies in this section.

4.1 Model

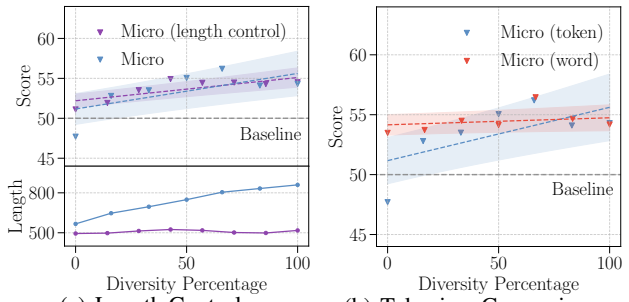
Building on our observations with Llama-2-7B, we conduct a cross-model analysis to examine whether the performance advantages of microscopic response diversity generalize across different model scales and architectures. We select two contrasting models for this investigation: the larger Llama-2-70B and the more recent Llama-3-8B. For each architecture, we construct two series of datasets of 10K samples through the microscopic strategy on response using their respective tokenizers, and train the models on these datasets. To enable meaningful comparison, we establish two baselines with either model respectively using randomly sampled datasets of equivalent size (10K samples) without diversity optimization and conduct the comparison.

As evidenced in Figure 5, both Llama-2-70B and Llama-3-8B mirror the performance pattern of Llama-2-7B. More strikingly, Llama 3-8B demonstrates an amplified correlation slope, suggesting heightened sensitivity to token-level diversity variations. While future works are required to investigate the deep reasons, we posit this difference stems from Llama 3's implementation of the tiktoken tokenizer, which enlarges the vocabulary size from Llama-2's 32K to 128K and the covering range of mid band tokens. Overall, these results confirm a robust and consistent positive correlation across different model sizes and architectures.

4.2 Length Control

As discussed in Section 3.2, response length exhibits a strong correlation with model performance scores. To investigate whether this relationship is causal, we conduct experiments controlling the response length to be identical. This ablation study uses the microscopic strategy on responses. Specifically, we construct length-controlled datasets by sampling from sub-datasets of varying lengths and selecting samples within a similar length range. Models are subsequently fine-tuned on these datasets and evaluated across diversity scales as presented in the proposed framework.

Figure 6 (a) reveals that while we restricted response lengths to approximately 500 tokens across all datasets, the relationship between diversity scale and model performance remains unchanged compared to the uncontrolled baseline.



(a) Length Control (b) Tokenizer Comparison

Figure 6: Ablation study on microscopic diversity: Figure (a) compares the performance under length control. The subplot below illustrates the average token length of responses across the datasets. Figure (b) compares word-based tokenization. The abscissa indicates the microscopic diversity percentage from the response perspective for the 10K dataset.

This persistent correlation suggests that response length may serve as an incidental covariate rather than the primary factor driving performance improvements.

4.3 Tokenizer

Given the crucial role of tokenization in the microscopic strategy, we investigate its influence by comparing the default tokenization scheme with word segmentation. The default tokenization aligns with the LLM to be fine-tuned while the word segmentation is a common approach that can be used for any LLM without any pre-processing step. We thus only replace the tokenize step by the word segmentation in the microscopic strategy, and construct a new series of datasets (of 10K samples) to train the model.

As shown in Figure 6 (b), word-based tokenization shows no clear positive correlation with model performance. This suggests that the tokenization scheme provided by the tokenizer is essential for preserving microscopic diversity, as it segments information in a way that matches input the model received during training. Word-based segmentation may disrupt this correspondence, leading to suboptimal performance.

4.4 Additional Ablation Studies and Discussions

Three additional studies are provided in Appendix C to further confirm our findings in various settings. (1) With original responses (which are of lower data quality), (2) On more benchmarks (MMLU (Hendrycks et al. 2021) and LiveBench (White et al. 2025)) and (3) with mismatched tokenizer to control the microscopic dataset diversity.

While prior research has predominantly emphasized the instruction diversity that implicitly leads to diverse responses and consequently diverse token distributions, our experiments reveal a more impactful approach. Through rigorous comparative analysis, diversity metric evaluations, and ablation studies, we conclude that the microscopic strategy on response is a more effective and direct way. We therefore encourage the research community to prioritize the response diversification alongside the instruction for SFT dataset construction.

5 Related Work

Recent advances in LLMs have underscored the importance of dataset diversity and token-level analysis. This section briefly overviews related work in these areas.

Dataset Diversity Approach. The assessment of dataset diversity has largely followed two paradigms. The first is a quantitative approach using syntactic-level features like distinct-n (Li et al. 2016) for n-gram uniqueness, the gzip compression ratio (Song et al. 2024) for redundancy, ROUGE-L for sequence overlap, and Self-BLEU (Zhu et al. 2018) for internal diversity. While intuitive, these metrics may not fully capture deeper contextual relationships. The second, a semantic approach enabled by modern language models (Vaswani et al. 2017), analyzes dataset structure more profoundly. This began with using BERT (Devlin et al. 2019) embeddings for vector distance measurement (Shaib et al. 2024; Liu et al. 2024; Wang et al. 2025a; Liu, Karbasi, and Rekatsinas 2024; Yang et al. 2025) and clustering (Grootendorst 2022; Ge et al. 2024b), and has progressed to using large generative models (Achiam et al. 2023) for sophisticated analyses like automated data tagging (Lu et al. 2023; Yang et al. 2024; Dubey et al. 2024) and agent-based cluster identification (Chen et al. 2024; Bai et al. 2025), thus providing a more semantic evaluation.

Token-Level Analysis in Language Models. Token-level analysis has been studied in LLMs, with numerous approaches exploring the relationship between tokens and various aspects of LLM training and performance (Zhong et al. 2023; Land and Bartolo 2024; Wang et al. 2025b). For instance, Lin et al. focus on training strategies that prioritize important tokens to enhance model efficiency. Madsen et al. observed variations in training time and gradient descent loss associated with processing different tokens, using these metrics to infer token importance or influence. Li et al. explored the use of token matching techniques to improve training efficiency and performance. While these studies provide valuable insights into token-level effects on training, the analysis of token diversity within the context of SFT datasets remains relatively unexplored.

6 Conclusion

In this study, we first propose a taxonomy to categorize the approaches for dataset diversity according to the grain level (macro-, meso- and microscopic) and the SFT datasets component (instructions and responses). Through extensive experiments, we systematically compare their impact on model performance and demonstrate that our proposed microscopic strategy on responses exhibits the strongest correlation between model performance and diversity degree while achieving superior performance at maximum diversity compared to other strategies. Detailed analyses and ablation studies confirm the effectiveness of the proposed microscopic strategy on responses. We also tested multiple diversity metrics and suggest that the information entropy may be a good estimator of dataset diversity, which also aligns with the optimization direction of the microscopic strategy. We therefore call upon the research community to prioritize the response diversification alongside the instruction for SFT dataset construction.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (U24A20264, 62273041), Open Projects of the Institute of Systems Science, Beijing Wuzi University (BWUISS11).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arora, S.; Liang, Y.; and Ma, T. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *Proceedings of the International Conference on Learning Representations*.

Bai, T.; Yang, L.; Wong, Z. H.; Sun, F.; Zhuang, X.; Peng, J.; Zhang, C.; Wu, L.; Jiantao, Q.; Zhang, W.; Yuan, B.; and He, C. 2025. Efficient Pretraining Data Selection for Language Models via Multi-Actor Collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9465–9491.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Bukharin, A.; Li, S.; Wang, Z.; Yang, J.; Yin, B.; Li, X.; Zhang, C.; Zhao, T.; and Jiang, H. 2024. Data Diversity Matters for Robust Instruction Tuning. In *Findings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3411–3425.

Chen, H.; Waheed, A.; Li, X.; Wang, Y.; Wang, J.; Raj, B.; and Abdin, M. I. 2024. On the Diversity of Synthetic Data and its Impact on Training Large Language Models. *arXiv preprint arXiv:2410.15226*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Du, Q.; Zong, C.; and Zhang, J. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ge, T.; Chan, X.; Wang, X.; Yu, D.; Mi, H.; and Yu, D. 2024a. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Ge, Y.; Liu, Y.; Hu, C.; Meng, W.; Tao, S.; Zhao, X.; Xia, M.; Li, Z.; Chen, B.; Yang, H.; Li, B.; Xiao, T.; and Zhu, J. 2024b. Clustering and Ranking: Diversity-preserved Instruction Selection through Expert-aligned Quality Estimation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 464–478.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the International Conference on Learning Representations*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Land, S.; and Bartolo, M. 2024. Fishing for Magikarp: Automatically Detecting Under-trained Tokens in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11631–11646.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119.

Li, T.; Chiang, W.-L.; Frick, E.; Dunlap, L.; Wu, T.; Zhu, B.; Gonzalez, J. E.; and Stoica, I. 2025. From Crowdsourced Data to High-quality Benchmarks: Arena-Hard and Benchbuilder Pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval. 2025-07-26.

Li, Z.; Chen, C.; Xu, T.; Qin, Z.; Xiao, J.; Sun, R.; and Luo, Z.-Q. 2024. Entropic Distribution Matching for Supervised Fine-tuning of LLMs: Less Overfitting and Better Diversity. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Lin, Z.; Gou, Z.; Gong, Y.; Liu, X.; yelong shen; Xu, R.; Lin, C.; Yang, Y.; Jiao, J.; Duan, N.; and Chen, W. 2024. Not All Tokens Are What You Need for Pretraining. In *Proceedings of the thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *Proceedings of the International Conference on Learning Representations*.

Liu, Z.; Karbasi, A.; and Rekatsinas, T. 2024. TSDS: Data Selection for Task-Specific Model Finetuning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Lu, K.; Yuan, H.; Yuan, Z.; Lin, R.; Lin, J.; Tan, C.; Zhou, C.; and Zhou, J. 2023. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *Proceedings of the International Conference on Learning Representations*.

Madsen, A.; Meade, N.; Adlakha, V.; and Reddy, S. 2022. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. In *Findings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1731–1751.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11): 205.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35.

Padmakumar, V.; and He, H. 2024. Does Writing with Language Models Reduce Content Diversity? In *Proceedings of the International Conference on Learning Representations*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Shaib, C.; Barrow, J.; Sun, J.; Siu, A. F.; Wallace, B. C.; and Nenkova, A. 2024. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*.

Song, F.; Yu, B.; Lang, H.; Yu, H.; Huang, F.; Wang, H.; and Li, Y. 2024. Scaling Data Diversity for Fine-Tuning Language Models in Human Alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 14358–14369.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, P.; Shen, Y.; Guo, Z.; Stallone, M.; Kim, Y.; Golland, P.; and Panda, R. 2025a. Diversity Measurement and Subset Selection for Instruction Tuning Datasets. In *Proceedings of the International Conference on Learning Representations Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Wang, S.; Yu, L.; Gao, C.; Zheng, C.; Liu, S.; Lu, R.; Dang, K.; Chen, X.; Yang, J.; Zhang, Z.; et al. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13484–13508.

Wang, Z.; Bukharin, A.; Delalleau, O.; Egert, D.; Shen, G.; Zeng, J.; Kuchaiev, O.; and Dong, Y. 2024. HelpSteer2-Preference: Complementing Ratings with Preferences. In *Proceedings of the International Conference on Learning Representations*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

White, C.; Dooley, S.; Roberts, M.; Pal, A.; Feuer, B.; Jain, S.; Shwartz-Ziv, R.; Jain, N.; Saifullah, K.; Dey, S.; Shubh-Agrawal; Sandha, S. S.; Naidu, S. V.; Hegde, C.; LeCun, Y.; Goldstein, T.; Neiswanger, W.; and Goldblum, M. 2025. LiveBench: A Challenging, Contamination-Free LLM Benchmark. In *Proceedings of the International Conference on Learning Representations*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, Y.; Nan, Y.; Ye, J.; Dou, S.; Wang, X.; Li, S.; Lv, H.; Gui, T.; Zhang, Q.; and Huang, X. 2025. Measuring Data Diversity for Instruction Tuning: A Systematic Analysis and A Reliable Metric. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 18530–18549.

Zhao, D.; Andrews, J.; Papakyriakopoulos, O.; and Xiang, A. 2024a. Position: Measure Dataset Diversity, Don't Just Claim It. In *Proceedings of the 41st International Conference on Machine Learning*, 60644–60673.

Zhao, H.; Andriushchenko, M.; Croce, F.; and Flammarion, N. 2024b. Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning. In *Proceedings of the 41st International Conference on Machine Learning*.

Zhong, Q.; Ding, L.; Liu, J.; Liu, X.; Zhang, M.; Du, B.; and Tao, D. 2023. Revisiting Token Dropping Strategy in Efficient BERT Pretraining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 10391–10405.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Texygen: A benchmarking platform for text generation models. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1097–1100.