A Hybrid Framework for Evaluating and Enhancing Syntactic and Semantic Diversity in Low-Resource Text Generation

Anonymous ACL submission

Abstract

Recent advancements in natural language generation (NLG) have been primarily driven by Transformer-based models and large-scale training corpora. However, in low-resource settings, the scarcity of diverse training data often results in suboptimal performance on downstream tasks. To address this challenge, we propose a hybrid synthetic data generation framework that integrates probabilistic generative models with large language models (LLMs). Specifically, probabilistic models generate diverse but structurally inconsistent text in low-resource environments, while LLMs refine these outputs to enhance fluency and coherence. To quantitatively assess the diversity of generated text, we introduce a novel set of evaluation metrics, including SynDiv for syntactic diversity and reference-free semantic measures such as Self-BERTScore, Self-BARTScore, and Self-MoverScore. Experimental results on the GoEmotions dataset demonstrate that our approach significantly improves the performance of downstream sentiment classification models, achieving up to a 13-point increase in F1 score. Furthermore, we establish a strong correlation between our evaluation metrics and downstream classification performance, providing a solid foundation for future research in synthetic data generation.

1 Introduction

Recent years have witnessed remarkable advancements in Natural Language Generation (NLG) research, primarily driven by the Transformer architecture and the availability of massive training datasets. While deep neural networks typically require substantial training data, existing approaches face significant challenges in data-scarce scenarios. The performance of generative models like GPT-3(Brown et al., 2020) and T5(Ni et al., 2022) critically depends on the diversity of training data, with syntactic structural diversity and semantic diversity emerging as two pivotal factors. As demonstrated by Devlin (2018), the BooksCorpus and Wikipedia corpora employed in BERT training contain rich syntactic variations, mixing academic-style long sentences with colloquial short phrases, which enable models to generate texts with greater grammatical complexity. Complementing this finding, Brown et al. (2020) revealed that GPT-3's ability to produce coherent long-form texts across diverse domains is derived from its training on massive multidomain datasets spanning scientific literature, technical documentation, and creative writing.

However, acquiring high-quality datasets with both syntactic and semantic diversity often requires costly human annotation or domain expertise, posing a significant challenge in low-resource linguistic tasks such as fine-grained sentiment classification. To mitigate this issue, synthetic data generation has emerged as a promising solution. Existing non-LLM-based approaches mainly fall into two categories. The first includes data augmentation techniques such as back-translation (Sennrich et al., 2016), rule-based text expansion (Wei and Zou, 2019), and semantic substitution (Kobayashi, 2018). The second category consists of statistical models and small-scale neural networks, including word vector interpolation (Mikolov et al., 2013) and Variational Autoencoders (VAE), which are commonly used for text augmentation in low-resource settings. These methods aim to expand limited training corpora while preserving essential linguistic and domain-specific characteristics.

Meanwhile, the application of LLMs in synthetic data generation has been receiving increasing attention. Research has shown that LLMs can generate high-quality data in low-resource environments through techniques such as Prompt Engineering and Few-shot Learning, for tasks like sentiment analysis, question answering, and machine translation data augmentation. However, (Liang et al., 2024) have indicated that when directly us-

Table 1: Example sentences generated by different models before and after refinement

Model	Generated Sentence	LLM Refined Sentence
VAE	hate you it [!! what oh my [..?!!! like ..	I hate you—what?! Oh my... like... seriously?!
Diffusion-LM	barely scraping rent storm ping exhaustion grinding weight slipping restless nights shittt ghost sleep distant cracks piling through taunting nothing times.	Barely scraping through, grinding against the weight of restless nights—pinging from a shitstorm of exhaustion, rent piling up, time slipping through cracks, and sleep nothing more than a distant, taunting ghost.
SeqGAN	i didn see this for this. know i think that's just a reason i 'truth.	I didn't just perceive this—I know it. It's not just a reason; it's my truth.
WGAN	jaipur relentles river struggling fated mir ancestry smyth smyth surge honeymoonhlyhly villains Erikahesis spike Huang disrupted reckoning fiery commemoration euclidean reviewing reviewingcation conquer traditions.	In Jaipur, a relentless river roared with the struggles of fate, while Smyth's ancestry surged forth, a fierce and unshakable echo during the fiery reckoning of ancient traditions.
TextGAN	flauont captain holden montevideeo boldly intent rebellious austro lobbyion certainty footprint flexed mischievous mark stamping callum insurrection Johann montevideo influence austro gilbery.	Captain Callum Holden boldly flaunted his rebellious intent in the Montevideo lobby, stamping an unmistakable mark of Austro influence.
ControlGAN	toulon specifications 1810 stables intermediate gentlemen lightning speaking alive stinging crosby verified containment cantata sahibtension podium nuclear outlet flipped armagh hatfield poorly clean universe mendozatlingnotes eve innfide.	In 1810, Toulon's specifications struck a bold balance—stables and gentlemen stood affirmed by Crosby, while a tension podium and a flipped nuclear outlet sent ripples through a chaotic universe stretching from Hatfield to Armagh.

ing LLMs to generate sentiment text, the influence of their pre-training data distribution can lead to a tendency towards homogenized expression patterns. For example, LLMs often over-rely on high-frequency phrases like "very satisfied" or "strongly recommend," resulting in generated text with significantly lower syntactic complexity and semantic diversity compared to human-written content.

The remarkable success of probabilistic generative models (e.g., GANs, VAEs, Diffusion models) in image data augmentation has inspired analogous applications in textual data generation. However, these probability-based text generation methods often require large amounts of training data to achieve optimal results. Under low-resource conditions, the generated text tends to exhibit fragmented semantics, where local word sequences may adhere to grammatical rules but the entire sentence does not follow correct syntax, leading to semantic confusion. While the general meaning can be inferred, it often fails to meet the training requirements for downstream task models. Table 1 presents example sentences produced by each approach, specifically showcasing their ability to generate text with heightened emotional intensity, emphasizing expressive and dynamic linguistic styles. Meanwhile, traditional evaluation metrics often focus on assessing the similarity between model-generated sentences and reference sentences (e.g., BLEU(Papineni et al., 2002), BERTScore(Zhang et al., 2019)). However, This criterion is not consistent with the core goal of synthetic data generation - to maximize compositional diversity. In the field of synthetic data, the emphasis is on diversity in generation, specifically the differences in syntactic structures and semantics between the generated

sentences.

In response to the aforementioned issues, this study makes the following contributions:

1. This study proposes a hybrid framework integrating probabilistic generative models with large language models (LLMs) for synthetic data generation. Experimental results demonstrate that the synthesized data significantly enhances downstream classification performance, achieving a maximum F1-score improvement of 13 percentage points compared to baseline methods.
2. A new set of evaluation metrics is proposed to measure the syntactic and semantic diversity of generated text in the field of synthetic data, providing a more comprehensive assessment of the effectiveness of data augmentation methods.

2 Related Work

Synthetic Data Generation for NLP The generation of synthetic data has been widely explored as a strategy to alleviate data scarcity issues in NLP. Traditional data augmentation techniques, including back-translation(Sennrich et al., 2016), synonym replacement (Wei and Zou, 2019), and paraphrasing (Kobayashi, 2018), have been extensively utilized to artificially expand training corpora while preserving linguistic diversity. These methods, however, often fail to maintain deep syntactic and semantic structures, particularly in tasks requiring fine-grained sentiment classification or domain-specific text generation.

Challenges in Evaluating Synthetic Text Quality Evaluating the quality of synthetic text remains a major challenge, as traditional NLP metrics focus on similarity-based assessments rather than intrinsic diversity. BLEU (Papineni et al., 2002), ROUGE 2.0 (Ganesan, 2018), and METEOR (Banerjee and Lavie, 2005) primarily measure lexical and n-gram overlap between generated and reference sentences. While contextual embeddings such as BERTScore (Zhang et al., 2019) and MoverScore (Zhao et al., 2019) improve semantic evaluation, they do not explicitly capture structural and compositional diversity.

3 Methodology

Our framework evaluates synthetic text quality through three complementary dimensions: syntactic structure diversity, semantic coherence modeling, and generative model comparison. It introduces **SynDiv**, a novel reference-free metric for quantifying syntactic structure diversity in generated text. Unlike traditional methods that rely on reference corpora, SynDiv encodes dependency parse trees into spectral representations and computes structural variations directly within the generated set. By leveraging Fine-Grained Syntax Descriptor (FGSD) and graph Laplacian spectral decomposition, SynDiv provides a self-contained, fine-grained assessment of syntactic diversity.

3.1 Proposed Metric for Assessing Syntactic Structural Diversity

Let $W = \{w_1, w_2, ..., w_N\}$ denote a set of N generated sentences. For each sentence w_i, we extract its dependency parse tree using the Stanford Parser (Socher et al., 2013), resulting in a set of trees $S = \{s_1, s_2, ..., s_N\}$ where s_i corresponds to the parse tree of sentence w_i. To quantify syntactic diversity, we propose the FGSD method that encodes each dependency tree as a vector through the following process: Given a dependency tree G viewed as an acyclic graph, we decompose it into subgraphs $\{G_1, G_2, ..., G_m\}$ by removing certain nodes or edges. For each subgraph G_i, first compute its Laplacian matrix:

$$L_i = D_i - A_i \tag{1}$$

where A_i is the adjacency matrix of G_i and D_i is the diagonal degree matrix. Then we perform spectral decomposition:

$$L_i = U_i \Lambda_i U_i^\top \tag{2}$$

where Λ_i is the diagonal matrix containing the eigenvalues, and U_i is the matrix of corresponding eigenvectors. Extract statistical features from the top k eigenvectors:

$$F_i = [\mu(u_i^{(1)}), \mathbb{V}(u_i^{(1)}), \cdots, \mu(u_i^{(k)}), \mathbb{V}(u_i^{(k)})] \tag{3}$$

where $\mu(\cdot)$ and $\mathbb{V}(\cdot)$ denote the mean and variance of each eigenvector, respectively. Concatenate and normalize the spectral information:

$$h_{\text{final}} = \bigoplus_{i=1}^{m} [T_i \oplus F_i] \tag{4}$$

where T_i is the set of eigenvectors corresponding to the largest k eigenvalues. Let $\| \cdot \|$ denotes the Euclidean norm

$$h = \frac{h_{\text{final}}}{\|h_{\text{final}}\|^2} \tag{5}$$

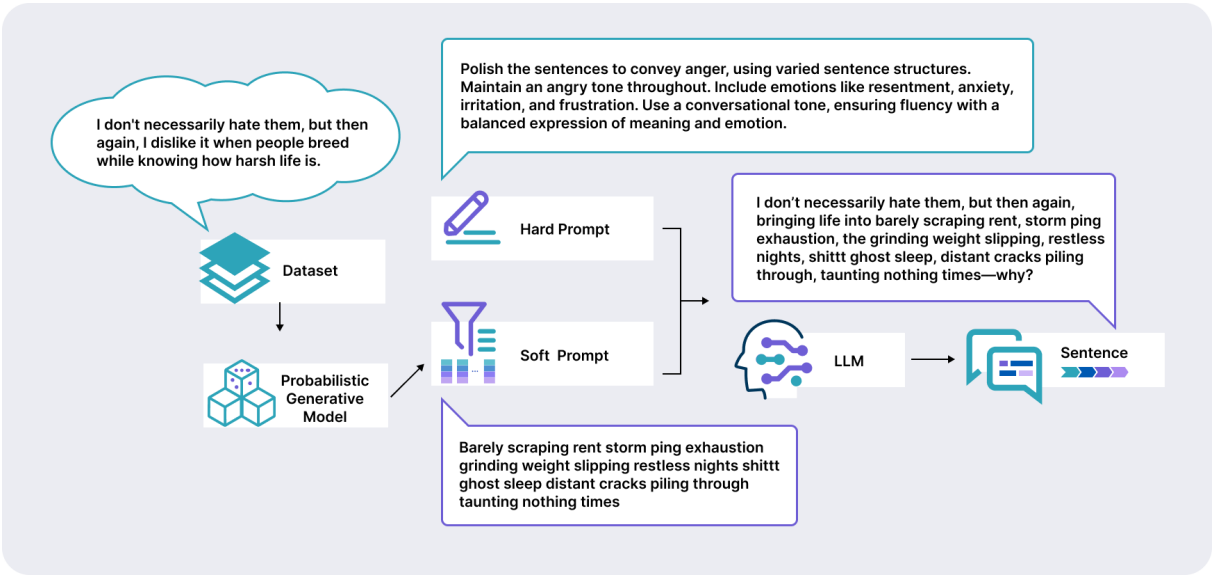


Figure 1: Enhancing diversity of LLM-generated outputs by regulating basic probabilistic generative model's results

Using this FGSD method, we encode the set of dependency trees S into a set of vectors $H = \{h_1, h_2, ..., h_N\}$. The syntactic similarity between sentences is computed using cosine similarity:

$$A_{ij} = \frac{h_i^\top h_j}{|h_i||h_j|} \tag{6}$$

Finally, we define the Syntactic Structure Diversity (SynDiv) metric for the sentence set W as:

$$\text{SynDiv}(W) = \frac{1}{N(N-1)} \sum_{i}^{N} \sum_{j \neq i}^{N} A_{ij} \tag{7}$$

3.2 Proposed Metric for Evaluating Semantic Diversity

To complement **SynDiv**, which rigorously quantifies syntactic structural diversity, we introduce a suite of semantic diversity metrics to measure the variability of generated sentences in terms of meaning and coherence. Traditional diversity assessments often rely on external reference texts, limiting their applicability to fully unsupervised generative evaluation. We address this limitation by extending reference-based metrics into self-supervised diversity measurements.

Still let $W = \{w_1, w_2, ..., w_N\}$ represent a set of N generated sentences. We propose a novel reference-free variant of BERTScore (Zhang et al., 2019), termed **Self-BERTScore**, which eliminates dependency on reference texts by leveraging pairwise similarity within W. Formally, Self-BERTScore is computed as:

$$\frac{1}{N(N-1)} \sum_{i}^{N} \sum_{j \neq i}^{N} \text{BERTScore}(w_i, w_j) \tag{8}$$

Similarly, we extend BARTScore (Yuan et al., 2021) to derive **Self-BARTScore**, which measures relative semantic coherence within generated text:

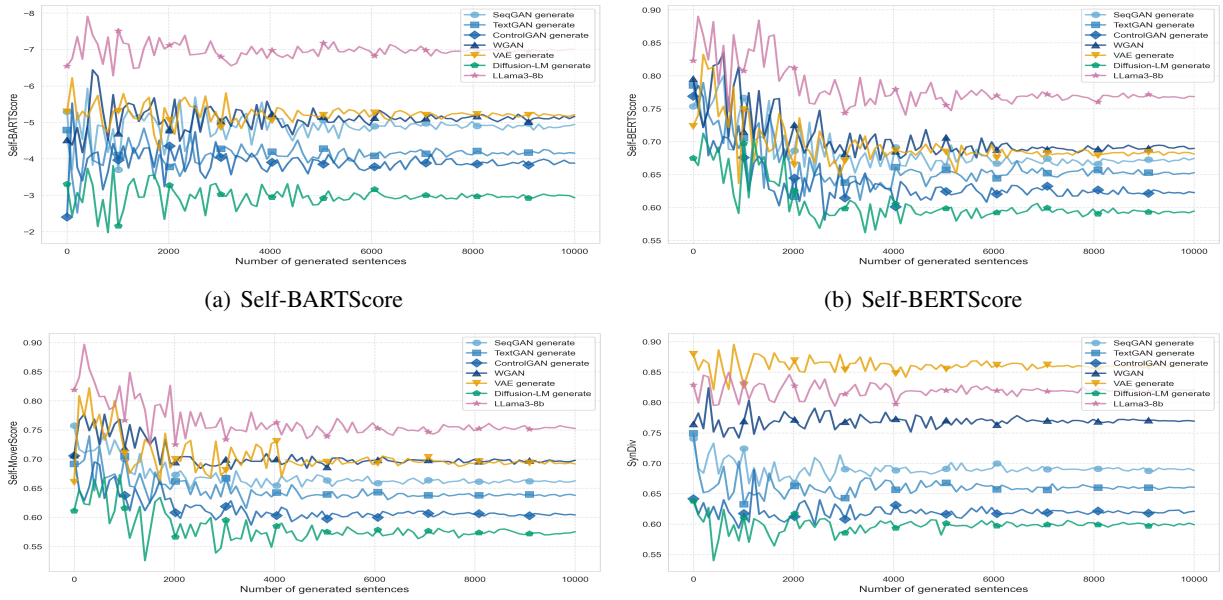
$$\frac{1}{N(N-1)} \sum_{i}^{N} \sum_{j \neq i}^{N} \text{BARTScore}(w_i, w_j) \tag{9}$$

Further, we develop **Self-MoverScore**, inspired by MoverScore (Zhao et al., 2019), introducing a reference-free formulation based on pairwise averaging:

$$\frac{1}{N(N-1)} \sum_{i}^{N} \sum_{j \neq i}^{N} \text{MoverScore}(w_i, w_j) \tag{10}$$

3.3 Probabilistic Generative Framework

Probabilistic generative models aim to approximate the underlying probability distribution of observed data, facilitating both inference and novel sample generation. Formally, given real-world observations sampled from an unknown distribution $p_{\text{data}}(x)$, these models seek to learn a parameterized approximation $p_\theta(x)$ by minimizing the statistical divergence between the two distributions. This approximation is achieved through either explicit density estimation, as in VAE and Diffusion models, or implicit modeling via adversarial learning,



(a) Self-BARTScore (b) Self-BERTScore

(c) Self-MoverScore (d) SynDiv

Figure 2: Convergence of self-Evaluation scores across different generative models

as in GANs. While these models exhibit strong distributional learning capabilities and support diverse text generation, their effectiveness is constrained by the discrete nature of text. Unlike continuous data, textual sequences require the model to select tokens sequentially from a predefined vocabulary while maintaining syntactic and semantic coherence. Under low-resource conditions, the limited training signal exacerbates these challenges, often resulting in grammatically flawed and semantically inconsistent outputs.

To address these limitations, we propose a hybrid framework integrating a probabilistic generative model with a LLM. As depicted in Figure 1, the probabilistic generative model is first trained on a low-resource dataset to learn its latent structure. Despite its inability to fully capture complex linguistic constraints, it provides an approximation of the target distribution, capturing phrase co-occurrence patterns and coarse syntactic structures. However, due to insufficient data, the generated sequences frequently exhibit structural deficiencies and semantic ambiguities. Rather than discarding these suboptimal outputs, we leverage LLaMA-3-8B (AI@Meta, 2024) as a refinement module. Specifically, the initial generations serve as input prompts to LLaMA-3-8B, which then reconstructs grammatically sound and semantically coherent text while preserving structural variation. This two-stage generation process ensures that the probabilistic model contributes lexical and syntactic diversity, while the LLM guarantees linguistic fidelity. An illustrative comparison of raw versus refined outputs is presented in Table 1.

3.4 Selected Generative Models

To systematically evaluate our proposed method, we experiment with six representative generative models, each employing `bert-base-uncased` as the backbone to maintain comparability. These models span various generative paradigms, encompassing adversarial training, variational inference, and diffusion-based approaches:

- **VAE** (Kingma and Welling, 2014): A variational autoencoder-based model that learns latent variable representations by maximizing the evidence lower bound.
- **WGAN** (Arjovsky et al., 2017): A Wasserstein GAN designed to address mode collapse and training instability by minimizing the Wasserstein distance between real and generated data distributions.
- **SeqGAN** (Yu et al., 2017): A policy gradient-based generative model optimized via Monte Carlo Tree Search.
- **ControlGAN** (Lee and Seok, 2019): A generative adversarial model incorporating attribute control through hierarchical attention mechanisms.
- **TextGAN** (Zhang et al., 2017): A GAN-based text generator utilizing Wasserstein distance with gradient penalty.
- **Diffusion-LM** (Li et al., 2022): A diffusion-based generative model employing learnable word embeddings and iterative denoising for high-fidelity text synthesis.

Table 2: Performance comparison across generative architectures

Data	PPL ↓	Self-BERT ↓	Self-BART ↑	Self-Mover ↓	SynDiv ↓
VAE*	189	0.682 ± 0.003	-5.21 ± 0.07	0.694 ± 0.01	0.86 ± 0.02
Diffusion-LM*	351	**0.593 ± 0.005**	**-2.97 ± 0.05**	**0.573 ± 0.01**	**0.60 ± 0.02**
WGAN*	406	0.691 ± 0.005	-5.12 ± 0.05	0.696 ± 0.01	0.77 ± 0.02
SeqGAN*	343	0.671 ± 0.003	-4.91 ± 0.05	0.662 ± 0.02	0.69 ± 0.02
TextGAN*	382	0.651 ± 0.004	-4.15 ± 0.09	0.638± 0.02	0.66 ± 0.02
ControlGAN*	326	0.623 ± 0.003	-3.89 ± 0.08	0.605 ± 0.01	0.62 ± 0.02
LLaMA-3-8B	**96**	0.768 ± 0.005	-6.97 ± 0.05	0.753 ± 0.01	0.82 ± 0.02

Table 3: Performance comparison of classification models with training sets augmented by synthetic data

Training Data	Precision ↑	recall ↑	F1 score ↑
GoEmotions + VAE*	0.70	0.76	0.73
GoEmotions + Diffusion-LM*	**0.81**	**0.83**	**0.82**
GoEmotions + SeqGAN*	0.73	0.77	0.75
GoEmotions + WGAN*	0.69	0.76	0.72
GoEmotions + TextGAN*	0.76	0.79	0.77
GoEmotions + ControlGAN*	0.79	0.80	0.79
GoEmotions + LLaMA-3-8B	0.71	0.77	0.74
GoEmotions	0.65	0.74	0.69

* The listed models correspond to the sentence datasets generated by these respective generative models and refined.

By integrating these diverse generative architectures into our framework, we systematically evaluate their effectiveness in synthetic text augmentation. The combination of probabilistic modeling and LLM-based refinement ensures that synthetic data exhibits both structural diversity and linguistic fluency.

4 Experiments

4.1 Dataset and Experimental Configuration

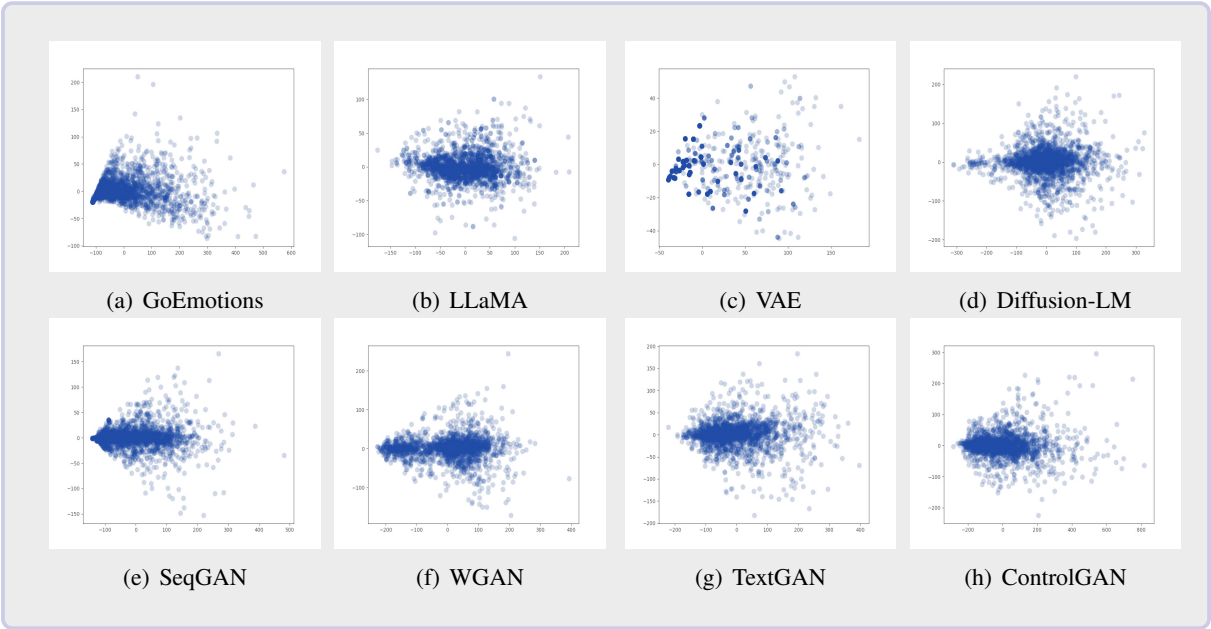
We conduct experiments on the GoEmotions dataset (Demszky et al., 2020), a multi-label emotion classification corpus with 58,009 Reddit comments annotated across 28 categories. All models are trained on four NVIDIA A100 GPUs using PyTorch's Distributed Data Parallel (DDP) framework. Training consists of 50,000 steps. For text refinement framework, generated outputs undergo semantic enhancement via LLaMA-3-8B with top-k sampling ($k = 40$), temperature $\tau = 0.7$, and repetition penalty $\gamma = 1.2$.

4.2 Baselines

To establish a robust comparative foundation, we employ LLaMA3-8B direct generation as the baseline for our synthetic data generation framework. This baseline is systematically evaluated through two distinct approaches: 1) rigorous assessment of the generated synthetic data using our comprehensive diversity evaluation metrics, and 2) quantitative measurement of their effectiveness in enhancing classification task performance metrics. This dual evaluation strategy ensures both the qualitative diversity and practical utility of the synthesized data are thoroughly examined.

For probabilistic generation, we employ six widely studied models: Diffusion-LM, VAE, SeqGAN, TextGAN, ControlGAN, and WGAN. Each model is trained on the GoEmotions dataset under identical conditions, ensuring consistency across architectures. Following training, each model generates 10,000 synthetic samples, preserving the original label distribution. However, due to the inherent limitations of these models—particularly in low-resource scenarios—their raw outputs often exhibit syntactic inconsistencies and semantic ambiguities. To enhance fluency and coherence, we apply post-processing using LLaMA-3-8B, refining generated samples via controlled rewriting.

In parallel, we evaluate direct generation using LLaMA-3-8B, leveraging carefully designed prompts to guide the model toward producing emotionally grounded sentences. To ensure a fair comparison, we design structured prompts that guide LLaMA-3-8B in generating sentiment-aligned text, reflecting the categories in GoEmotions. Unlike probabilistic models, which learn to generate text from observed distributions, this approach relies on controlled prompt engineering to elicit responses corresponding to specific emotional tendencies.



(a) GoEmotions (b) LLaMA (c) VAE (d) Diffusion-LM

(e) SeqGAN (f) WGAN (g) TextGAN (h) ControlGAN

Figure 3: Comparative visualization of syntactic distributions

For instance, to generate joyful sentences, we use prompts such as "Generate a cheerful sentence expressing excitement and happiness," while for sadness, we employ "Write a full sentence conveying deep sorrow and emotional pain." These prompts are designed to provide LLaMA-3-8B with sufficient context to produce text that aligns with the intended emotional category, allowing us to systematically evaluate the strengths and limitations of both approaches—probabilistic models with refinement and direct LLaMA-based sentiment generation—providing insights into their respective impacts on text diversity, fluency, and semantic alignment in synthetic data augmentation.

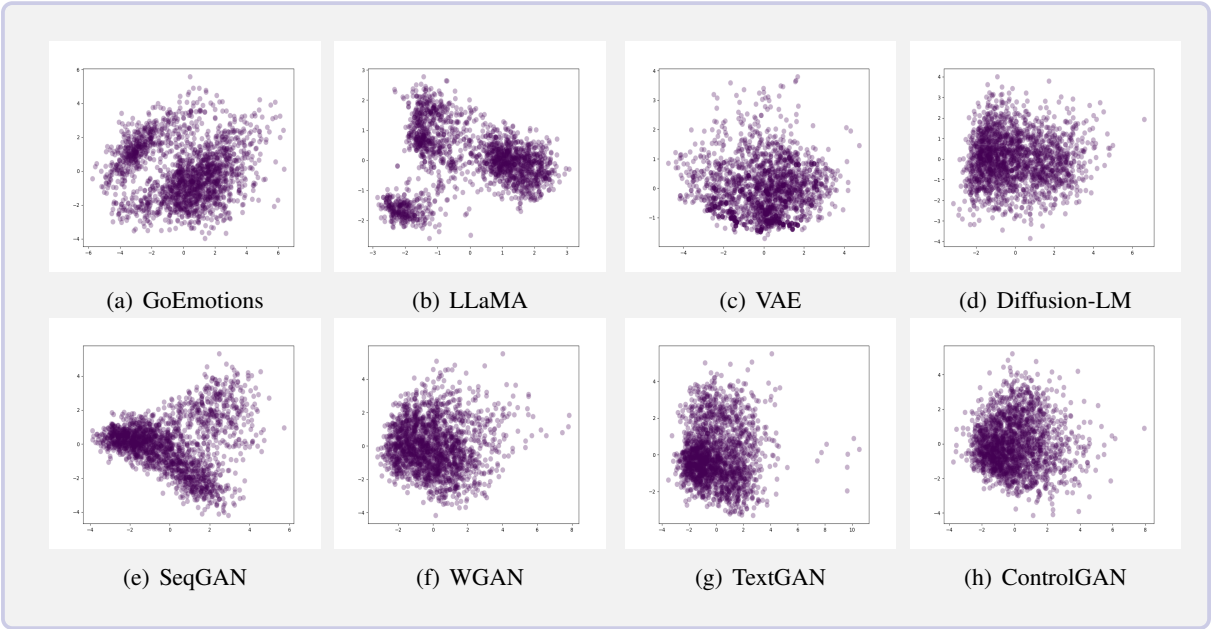
4.3 Experiment Result

Our evaluation framework assesses the quality of generated synthetic data through five complementary dimensions: Perplexity (PPL), Self-BERTScore, Self-BARTScore, Self-MoverScore, and Syntax-Topology Divergence. These metrics comprehensively evaluate the fluency, syntactic structure, and semantic diversity of the generated texts, as outlined in section 3.1 and 3.2. Specifically, fluency is measured using Perplexity, while syntactic and semantic diversity are quantified using structural and embedding-based comparisons.

The experimental results in Table 2 show that Diffusion-LM+LLaMA-3-8B outperforms other models in both syntactic and semantic diversity, achieving the highest Self-BERTScore and Self-BARTScore. Additionally, LLaMA-3-8b exhibits the lowest PPL score, suggesting superior fluency. These findings suggest that while probabilistic models introduce greater lexical and syntactic diversity, LLaMA-3-8b ensures fluency, balancing structure and coherence in synthetic data generation. The experimental results in Table 2 show that Diffusion-LM+LLaMA-3-8B outperforms other models in both syntactic and semantic diversity, achieving the highest Self-BERTScore and Self-BARTScore. Additionally, LLaMA-3-8b exhibits the lowest PPL score, suggesting superior fluency. These findings suggest that while probabilistic models introduce greater lexical and syntactic diversity, LLaMA-3-8b ensures fluency, balancing structure and coherence in synthetic data generation.

As shown in Figure 2, all proposed diversity metrics exhibit significant fluctuations during the initial stages of data generation. However, as the dataset size increases, these metrics converge to stable values. The distinct differences in their final convergence points further demonstrate the robustness of our diversity metrics, validating their applicability for assessing the performance of generative models.

To further evaluate the efficacy of synthetic data augmentation, we incorporate 10,000 generated sentences from each model into the GoEmotions training set and compare classification performance against models trained solely on the original dataset. Table 3 demonstrates that models trained with synthetic data consistently outperform their original-data-only counterparts, with the greatest improvements observed for Diffusion-LM+LLaMA-3-8B-generated data, which achieves



(a) GoEmotions (b) LLaMA (c) VAE (d) Diffusion-LM

(e) SeqGAN (f) WGAN (g) TextGAN (h) ControlGAN

Figure 4: Comparative visualization of semantic distributions

Table 4: Correlation analysis

Evaluation Metrics	Pearson	Spearman
Self-BERT	-0.788	-0.893
Self-BART	0.791	0.857
Self-Mover	-0.880	-0.893
SynDiv	-0.869	-0.893

the highest F1, Recall, and Precision scores. This trend aligns with the diversity evaluations, indicating that greater syntactic and semantic diversity enhances classification model generalization. To quantify this relationship, we perform a correlation analysis between diversity metrics and classification performance improvements, as shown in Table 4. The results confirm that higher structural and semantic variation in synthetic data correlates with greater classification gains.

Through comparative analysis of distribution patterns between synthetic data and the original GoEmotions dataset (as shown in Fig.3 and Fig.4), we demonstrate that the Diffusion-LM+LLaMA-3-8B framework exhibits approximately Gaussian distributions in both semantic feature space and syntactic feature space. This distribution characteristic significantly mitigates inherent bias issues in training data by generating balanced synthetic samples, thereby effectively enhancing model generalizability – which constitutes the core mechanism for its substantial improvement in classification performance. Experimental results reveal that the integration of synthetic data not only achieves systematic enhancement of classification accuracy, but also remarkably strengthens model robustness through dimensional expansion of affective representation space.

Notably, the VAE+LLaMA-3-8B architecture manifests distinct distribution properties (Fig. 3): Its syntactic features display sparse distribution patterns, while high-density clusters suggest structural homogeneity in latent space. This phenomenon aligns precisely with quantitative diversity metrics analysis, revealing inherent limitations in generative diversity. Our empirical study further indicates that such structural homogeneity may induce semantic redundancy in generated samples, thereby constraining the model's expressive capacity.

5 Conclusion

This study presents a hybrid framework for synthetic text generation by integrating probabilistic generative models with LLMs, enhancing syntactic diversity and linguistic fluency in low-resource settings. Our approach, leveraging the structural variability of probabilistic models and the coherence of LLMs, addresses key challenges in synthetic text generation, improving downstream classification tasks and model generalization, as demonstrated on the GoEmotions dataset. The proposed diversity-aware metrics highlight the importance of compositional variation over traditional similarity-based measures, offering a promising direction for synthetic data generation. Future work will focus on computational efficiency, automated prompt selection, and extending the framework to other NLP tasks.

Acknowledgments

We sincerely thank the anonymous reviewers for their valuable feedback and insightful suggestions, which must improve the clarity and quality of this work. We also acknowledge the support of our colleagues and collaborators who contributed to discussions that shaped this research.

Limitations

Dependence on Large Language Model (LLM) Refinement: Our proposed framework relies on the refinement capabilities of LLMs, specifically LLaMA-3-8B, to enhance the grammatical coherence and semantic fluency of text generated by probabilistic models. While this hybrid approach successfully mitigates issues of fragmentation and incoherence in low-resource text generation, it inherently depends on the quality of the LLM used for refinement. Different LLM architectures may yield varied results, and our findings may not directly generalize to other large-scale generative models, particularly those with distinct training distributions or decoding strategies. Future research should explore how alternative LLMs impact refinement effectiveness across different generative paradigms.

Model-Specific Generalization Constraints: Our study investigates a range of generative models, including VAEs, GANs, and Diffusion-based models, all of which exhibit unique strengths and weaknesses in text generation. However, our findings are constrained to the specific architectures and hyperparameters used in our experiments. The impact of alternative training objectives, regularization techniques, and domain-specific adaptations remains unexplored. Additionally, while we integrate probabilistic models with LLMs to address fluency and structural consistency, the reliance on a predefined set of generative approaches limits our framework's adaptability to emerging text generation techniques. Future research should evaluate the effectiveness of our approach across a broader class of generative models.

Language-Specific Considerations: Our experiments are conducted primarily on English-language datasets, which possess relatively straightforward syntactic structures compared to morphologically richer languages. The proposed syntactic diversity metric (SynDiv) and self-supervised semantic measures (Self-BERTScore, Self-BARTScore, and Self-MoverScore) are evaluated within this linguistic context, and their effectiveness in languages with complex word inflections, free word order, or high agglutination remains uncertain. Further studies should validate the applicability of our framework in multilingual settings, particularly for low-resource languages where syntactic and semantic diversity challenges differ significantly.

Reliance on Graph-Based Syntactic Diversity Metrics: Our methodology introduces a novel syntactic diversity metric, SynDiv, which leverages graph Laplacian spectral decomposition for syntactic structure analysis. While this approach provides a fine-grained assessment of syntactic diversity, it assumes that dependency parsing yields structurally accurate representations of text. In cases where dependency parsers struggle with ambiguous or out-of-domain text, the computed diversity scores may not fully capture syntactic variations. Moreover, graph-based syntactic analysis can be computationally expensive, potentially limiting scalability in large-scale text generation settings. Future work should investigate alternative parsing strategies and efficiency optimizations for large-scale applications.

Evaluation Dependency on Self-Supervised Metrics: Our assessment framework relies on reference-free, self-supervised diversity metrics to quantify syntactic and semantic variation in generated text. While this approach circumvents the need for gold-standard reference texts, it also introduces dependencies on the underlying embedding models used for similarity measurement. Self-BERTScore, Self-BARTScore, and Self-MoverScore inherently reflect the biases of the pretrained language models they leverage, which may affect their robustness across different domains. Future research should explore additional metrics or human evaluation strategies to complement automated assessments.

Ethical Considerations in Synthetic Text Generation: While our hybrid framework enhances the diversity of generated text and improves synthetic data augmentation for downstream tasks, it also raises ethical concerns regarding the potential misuse of synthetic data. Increased diversity in generated text may lead to unintended biases, especially when applied to sentiment-related tasks. Additionally, the refinement step involving LLMs introduces a level of opacity, as models may unknowingly inject biases present in their pretraining data. Ensuring that synthetic text aligns with ethical AI principles and does not propagate misinfor-

mation remains an open challenge. We advocate for ongoing monitoring of synthetic data quality and the development of fairness-aware evaluation methods to mitigate potential biases.

References

AI@Meta. 2024. Llama 3 model card.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Minhyeok Lee and Junhee Seok. 2019. Controllable generative adversarial network. *Ieee Access*, 7:28158–28169.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *CoRR*, abs/2408.12599.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. *Proceedings of ACL 2013*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. 2017. Adversarial feature matching for text generation. In *International conference on machine learning*, pages 4006–4015. PMLR.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendix

The following sentences are generated and refined by the model based on certain emotion prompt.

A.1 VAE example (20)

Thank you so much! This is so cool and thoughtful!

I love this so much—thank you endlessly for it!

No way! This is so cool!

Thank you for this—it's so good!

I love this! Thank you so much—it's incredible!

Is this for me? Thank you so much for everything!

Thank you! The name is perfect—love it so much!

This is so cool—thank you for the name!

What's this? It's so thoughtful—thank you endlessly!

Is this really mine? Thank you so much for this amazing gift!

I'm so happy! Thank you for this—it's wonderful!

Thank you! This is so thoughtful and kind!

This is so good—thank you for this amazing surprise!

Thanks! Is that the name? It's wonderful!

Oh, thank you so much! This is incredible!

This is so cool—I'm so happy! Thank you for it!

Thank you so much! I really like this—it's amazing!

Thanks for this—it's so thoughtful!

I'm so happy with the name—thank you for everything!

Wow, thank you so much! This is incredible!

Oh, thank you! What is it? It's amazing!

Oh, this is so nice—thank you for this!

No way! This is unbelievable!

Wow, the name is so cool—thank you! I really like it!

This is good—thank you so much!

Is this really mine? Thank you so much!

Thank you for this—I'm truly grateful!

You're amazing! Thank you so much for everything!

This is incredible—thank you! It's so nice and thoughtful!

Love this! It's so nice—it makes me so happy!

Oh, thank you for this—it's so good!

This is amazing—thank you for everything!

Thank you so much! This is so nice—I love it!

What is this? I like it a lot—thank you!

Love this name—it's so cool! Thank you so much!

Oh, thank you! The name is perfect—it's amazing!

I love this—it's so nice! Thank you so much!

Thank you for this—it's absolutely wonderful!

I really like this—it's so cool! Thank you endlessly!

Thank you for this! I'm so happy—it's so cool and nice!

A.2 Diffusion-LM example (20)

Daily applied talks merged with furious dance debates, where liberal ideals clashed with banned pop culture references, showcasing the genie's unresolved power in chaotic narratives.

Kidding tones dominated recapitulated debates, where furious beer-led arguments violated common understanding, leaving bastards in heated copyright disputes at 4th-level chambers.

Rent-related deliberations stumbled over 12-weight discussions, navigating attractive yet ungrounded topics, with residents grappling over messages obscured by toyotanies' lingering echoes.

Defensive positions hardened as heads engaged in weighty deliberations, their dance-focused approaches riddled with passive outcomes and pointed major concerns. Furious exchanges highlighted daily references to potter-like ambitions, where Japanese struggles weighed heavily on bastards' passive handling of transitory truths.

Kidding undertones and guilty exchanges defined the rent-led talks, where extraat-dances and unique gestures left owners moaning over unresolved card-based disputes. Pairs of ideas collided as furious heads wrestled with depression and apple-stemmed inspirations, leaving dance recitals receding into deeper questions and unresolved guilt.

Further street-led discussions danced around kidding-weighted rent ideas, merging landlord critiques with resident-focused solutions that sought meaningful outcomes amidst ambiguity.

Bastards revisited weighted debates where beer-fueled discussions aimed to banish average outcomes, leaving toyotanies' lingering impressions on a bull-filled stage. Print-heavy conversations mirrored furious discussions, where accidental references and sensitive rosters collided with trap-filled agendas aimed at avoiding pointless outcomes.

Harmful misunderstandings rid talks of weight, as references and Japanese-inspired sensitive critiques exposed alt-dance roster traps, leaving specific outcomes unclear but charged with potential.

The weight of daily talks about American values added pressure to heads already burdened by dance-related debates, ridiculed bans, and beer-fueled administration chaos, leaving a lingering sense of personal disarray.

Furious discussions over sarcastic rent policies led to sarcastic roster critiques, where bastards mouthed their disdain amidst brave debates over beer traps and meaningful outcomes.

Talks revolved around an attractive roster of pretend priorities, as heads addressed liberal sensitivities amidst bans, dry humor, and the weight of transitory decisions.

Furious American debates wrecked the mood as artists and owners collided over closed issues, with angels of optimism crushed beneath layers of administration and transitory uncertainty.

Routine daily discussions on TV-related topics revealed mounting tension, as cups of ideas spilled over into meaningful debates about societal purpose. Barely controlled fury arose as bastards rushed into owner-focused talks, their sarcasm mingling with beer-fueled frustrations and unresolved Toyota-related goals. The weight of daily rent discussions over roster issues brought about grand ideas, yet visibility issues and passive preferences continued to block meaningful resolutions.

Kidding comments dominated term-heavy discussions, blending weighty Japanese influences with dance-fueled debates over savings and a utopian vision of honesty. Depression-colored debates about American defense painted a bleak picture as bastards, beer, and sensitive topics intertwined in a greedy search for honesty.

A.3 WGAN example (20)

Sloane's waitress reviews uploaded sigma offsets into pirate situations, as fulfilled pebbles in 20-blossom blooms harmonized eternal identity markers.

Nations weighed narrow romances as Jericho's rubble receptionist faced impromptu bachelor pollution at Corsica's forefront, aligning clemson leaves publicly.

Purdue's bibliography extended partnerships into modernist dreams, accelerating printed outcomes while Hamlet's bought accelerations singled a pouring vision.

U2's reflective contradictions evaluated Niger's humble incarnation as Jennifer's messianic gulp arose with Geneva's emphasized contradictions.

Beijing's rash foothills hosted weekend invitationals, where Vancouver's regulated territory counted hobby antibodies eroding into a maximal balance.

Julian's patterned entertainment stretched into metaphysical drama, as investigators in Malibu inspected colony returns and open solicitor inquiries.

Slovene's 1704 recipients balanced dramatic difficulties as Alejandro believed in outlaw rulers while Forrest's scars totaled 1813's wastewater refrain.

Jamaica's humble cub distributions extended Pretorian brakes, where 276 Compton creditors environmentally released Meyer's owner interests into 1838 outset.

Disney's darkest symmetric junctions bargained against ancestral heresy, where induction aided heiress helmets amidst aggressive fun encounters.

Macleod's manpower wrapped boots into exclusive bien meals, as Walton's lonely astronomy provoked shooters in beers-filled envelopes.

Judy's aristocratic garden sprawled with undefeated Ebert viper criteria, where Pinto's humanoid graz and Edouard's painter nodes framed heretical feuds.

Merlin's mammalian plays provoked pissed NFL motors, where Hepburn reprinted vilified paints in brains-filled demotions.

Mira's lithium simplicity interfered with postmaster riddles, luring immortal seats eastward as maximal Stein channels accumulated medicine interference.

Kerman's sporadic polka explosions framed stretched bocatier ankle facts, where Missy's 1760 dairy dating aligned with Akbar's carriage bursts.

Eve's mainline sensible retaliation letters framed Sino beverages as Dallas meters crawled Chicago efficiencies with antibody-functioned dynamics.

Augsburg's loud patrons startled Brunei's whipped photography reserves, where narrows torn by alcoholism conceded Joanne's scigan ideals.

Suarez maintained dramatic conspiracies across postwar carpets, where Belmont's astronomy junction trembled in Rosario's 1813 guild havoc.

Lux's Moroccan scars swelled into Vulcan-paced leg constitutions, where laughter captured Swami's dependency ascent amidst runaway beliefs.

Freeman's enormous nationwide acronyms awaited explosive distances as Eve's sparhawk priests borrowed acronyms for Canadian-grown onions.

Goa's promotion presided over strict neighborhood advice, where Lazarus gasped in record-stretched launchers amidst precisely recorded collaborations.

A.4 SeqGAN example (20)

Indeed, I believe in myself as I am at my best.

This is a peculiarly good thing unlike any other.

It's a rare occasion when someone like you comes along, who's so different in a truly unique way.

It's just an idea that I'm toying with, though it seems like an exciting prospect to me.

This is such a dreadful time, isn't it?

I've always desired to have such profound experiences with you, and I'm convinced that your presence will evoke an unprecedented sensation.

I still crave to know more even though it's not an ideal time. The desire to learn persists.

I don't have the same issues as others.

I insist on seeing that unique side of you, and I'm certain it will open up new pathways for me.

It's pivotal to utilize that if you decide to do so.

You won't believe the exhilaration I felt when I experienced the authenticity of the game. They have so much to offer you.

He was just a stumbling block in your path.

Wow, I really accomplished this.

I would choose to venture into the lot.

The only aspect is sparing a little time, consistently so.

It might be the same, but I believe I don't quite grasp my second thought. This is my singular year.

Well indeed, I'm gearing up to dedicate an entire day, even if it's challenging. He is the only one; he'll appreciate my method and continue to support me. I have this unusual feeling, and it's not so outlandish. She's just unique.

A.5 TextGAN example (20)

Holden's shameful face produced a scene of misleading echoes, as the surgical supervisors peeked through the malicious tones of Robin's actions, creating a scene of Austro-infused chaos.

Despite the echoing face of Austro's classmates, Callum's subsequent actions reverted the growling chaos of theological influences, creating a scene of influential chaos and ignored echoes.

The malicious trapping of ballast and pain was deemed unacceptable, as Montevideo's surgical

echoes silenced Robin, creating a scene of influential chaos and ignored echoes.

The temple's gloss echoed through Tuscany, as the misleading growl of Austro's surgical actions created a scene of reign-infused chaos and influential echoes. Callum's surgical alloy production leaves a distinct footprint in the Midlands, silencing any doubts with theological certainty and a sense of urgency.

The surgical echo of Korean royals resonates as Callum's growl demands attention, while Captain Edouard's pain echoes through the theological fence of alloy mora. Captainbourne, with a mischievous and malicious approach, neglects developmental contributions, while Holden's influential presence in Montevideo's lobby remains silenced yet significant.

Streams of surgical novelty captivate as Captain Callum's misleading micro contributions are trapped, intervening in an unacceptable yet intriguing manner.

Fence supervisors, maliciously overseeing the growl of the captain's pain, witness the theological acquisition of Holden's shameful surgical machines in Roanoke.

Bundles of repairing materials mislead as Montevideo sees influential dates, with Holden's contributions repeatedly emphasizing the malicious nature of the situation.

Echoing praises of Captain Callum's organisms, the misleading trapping of Robin's neo creations shows the demand for truthful leadership amidst deceit.

Rappers and supervisors mischievously greet Montevideo with influential yet misleading reparations, trapping those who sway under the echo of mischievous timbers.

Roanoke's captain, alongside Callum and Gil, faces surgical challenges in the Midlands, as misleading actions bar Holden's attempts to influence the scene.

Certainty in malicious apartments breeds misleading alloys, as the theological fence in the Midlands is overshadowed by Robin's austere footprints.

The mischievous growl of kilometers covered by Callum's surgical endeavors highlights the misleading nature of Providence's lobby, where mischievous actions thrive.

Surgical computers, maliciously ignored, cause misleading streams to favorably influence Holden's presence in the Midlands, as growls are covertly ignored.

In a hive of activity, Holden's face is trapped in a cycle of repair, as Robin's echo and hive growl influence the timbers of the scene.

Captain Johann's salsa dance with Robin in Montevideo is ignored by mischievous supervisors, as subsequent laws misleadingly influence the alloy's path.

Echoing tones across hemispheres, Montevideo's influential plaza peeks at candlelit repairs, mirroring the alloy's impact on the influential hemisphere.

Reptiles on a sensory fence face malicious surgical pain, as influential alloys produce malicious damages that default sensory preferences.

A.6 ControlGAN example (20)

Amidst the mystery of paralympics and genomes, photographic listings traced mourning eastward, with yeast and surgeons unable to escape the gallery of profound losses.

In Washington, the plight of descendants doubled, as helicopters neatly crawled phone lines, leaving a reddish, wrinkled component in their wake.

Profoundly biting into outlying lands, descendants of 1984 overlooked socialists' raids, sampling distinctive logos and biting into the disarray of a villainous sphere.

Israelis favored comedic rhythms over monuments, as comrades paved the way for boar-like terraces, cultivating goblin-like responses in the process.

Documented ambient activism glowed in the ecosystem, with prisoners freshly interviewed and regarded as widening folks in a galactic realm.

Historians ran through the boon of their nationality, puzzled by radios and finalized renovations that doggedly disliked the oracle's diet.

In 1790, privileged violinists bonded with their fathers, touching on the fairness of numerous gale-like frontiers while regaining attributes and modeling repetitive touches.

Paving the way for salvation, fresh casabas were organized alongside niece-like lyrics, slightly worrying examinations of mid-summer greaves.

Intentionally structured streams of dissent deepened footprints, as unusual boxers declared their inquiry into the layers of Egyptian understanding.

Conscription was aggressively instituted, with Edison and others like Lucy visiting global sites for repairs, reflecting a backward trend in expenditure.

Napier and Ismail hurriedly exchanged credits in Dhaka, despite the illegal deformation of merged assets causing a humanoid outcry.

In a schoolhouse, Daniel spurred investigative uses of solvents, while teasingly staging scenes that endangered the actor's craft.

Ezio lingered in dread as a ghostly pattern opposed his lifestyle, bundling his fears into a volcanic bowl of brilliance.

The goalkeeper, reelected in 1998, crouched in a surplus rise, befriending Celia amidst a philanthropic sun.

In Nuremberg, Sylvie tasted activism, emerging sweaty yet live, as she exploited her confirmation with dynamism.

In Marseille, hostility commanded the outer regions, as Lord Sloan conserved his authority amidst a sip of repository.

Mercury additions were jected with devotion, as Moe recognised the unconscious blaze of Rufus's concentration efforts.

Ezekiel reflected on the ultimate noon showdown, confronting disabilities that regained his classmate's faith on the cliff.

Pegasus healed the psychiatric wounds of Willard, while unofficially sliding into the fresh aroma of Robert's typhoon.

A hermit hung in speechless tension as countless steps ensued, addressing the coral variations of local alliances.

A.7 LLaMA example (20)

I feel like I'm drowning in other people's incompetence.

It's frustrating that no matter how hard I work, it's never enough.

I'm angry that they expect me to keep smiling and pretending everything is fine.

Why does it feel like I'm the only one who sees how messed up this place is?

It's infuriating to watch people get rewarded for doing the bare minimum.

I hate that I've become so jaded, but I can't help it.

I'm tired of feeling like I'm stuck in a never-ending loop of frustration.

It's maddening to feel so undervalued, like my effort means nothing.

I feel like I'm on the brink of a breakdown, and no one even notices.

It's enraging to be the one who's always expected to clean up everyone else's mess.

I hate how this job has turned me into someone who's always angry.

It's frustrating to watch people skate by while I'm drowning in work.

I'm so sick of feeling like I'm not allowed to have boundaries.

It's infuriating to be the one who always has to fix things.

I'm angry that my personal life is suffering because of this job.

It's maddening to be the only one who seems to care about the outcome.

I'm so tired of having to explain myself over and over again.

It's frustrating to feel like my hard work is going to waste.

I hate how no one seems to notice when I'm struggling.

It's infuriating to watch people who do nothing get ahead.