

Tackling OOV: Masked Language Models vs Character Diffusion

Khondoker Ittehadul Islam (Matr. No: 7085810)

1 Motivation

Large Language Models (LLMs) have demonstrated remarkable success across a wide range of natural language processing (NLP) tasks, including text classification, generation, and reasoning. Despite these advancements, a persistent limitation remains in their ability to generalize effectively in the presence of noisy or out-of-vocabulary (OOV) inputs. This issue is particularly pronounced in informal domains such as social media, where spelling variations, abbreviations, and creative language use are prevalent.

Modern LLMs typically rely on Masked Language Modeling (MLM) objectives and subword tokenization techniques such as Byte-Pair Encoding (BPE) to mitigate the OOV problem. While BPE reduces the frequency of completely unseen tokens by decomposing words into subword units, it does not fully resolve challenges arising from orthographic variation or rare word distributions. Prior work has shown that even with subword modeling, performance can degrade significantly on low-frequency or noisy tokens, sometimes falling behind traditional feature-engineered approaches [IKIA21].

Motivated by this limitation, this work explores an alternative paradigm based on Character Diffusion Language Models (DLMs). Unlike MLMs, which learn to predict masked tokens, diffusion-based models iteratively corrupt and reconstruct sequences. In this study, we hypothesize that applying diffusion at the character level allows the model to better capture transformations between corrupted and clean textual forms. Specifically, by learning to denoise character-level perturbations, the model may implicitly learn robust representations of orthographic variation, thereby improving generalization in OOV scenarios.

To the best of our knowledge, the application of discrete character diffusion models for addressing OOV in downstream classification tasks remains underexplored. Therefore, this work aims to provide an initial empirical investigation into this direction. To facilitate reproducibility and further research, the implementation and experimental setup are publicly released¹.

2 Methodology

The central objective of this study is to compare the effectiveness of MLM and DLM paradigms under controlled conditions, particularly focusing on their behavior in OOV-rich environments. To ensure a fair comparison, both models are trained on the same corpus and share similar architectural configurations.

2.1 Models

2.1.1 Masked Language Model (MLM)

For the MLM baseline, we adopt a lightweight implementation based on nanoGPT². The model is trained using a standard token prediction objective, where a subset of tokens is masked and the model learns to reconstruct them based on surrounding context. This approach has been widely used in transformer-based architectures and serves as a strong baseline for contextual representation learning.

2.1.2 Diffusion Language Model (DLM)

For the diffusion-based approach, we utilize a discrete character-level diffusion model built upon nanoGPT. The implementation follows the framework described in [Kum25], incorporating the score-entropy loss proposed in [LME23]. Unlike MLMs, DLMs model a sequence of progressively corrupted inputs and learn to reverse this corruption process. By operating at the character level, the model is expected to capture fine-grained variations in spelling and structure, which are often the source of OOV challenges.

2.2 Data

2.2.1 Evaluation Task

The models are evaluated on the SemEval-2019 Task 6, Subtask A (OLID) [ZMN⁺19], which focuses on offensive language identification in tweets. This dataset is particularly suitable for OOV analysis due to its informal nature,

¹<https://github.com/KhondokerIslam/mlm-vs-dlm-on-ooov>

²<https://github.com/karpathy/nanoGPT>

Dataset Size (in char.)	1,188,765
Vocab Size	169
File Size	1MB

Table 1: Corpus Details which was used to train the foundational model

including frequent use of slang, misspellings, and domain-specific expressions. An important characteristic of this task is that offensive content often contains higher levels of lexical variation and noise compared to non-offensive text. This makes it an effective benchmark for assessing how well models can generalize across varying OOV distributions.

2.2.2 Training Corpus

Both models are trained on the OmniGEC Human Spell Corrected Reddit corpus [KRI25]. This dataset consists of user-generated comments paired with human-corrected versions, providing a natural bridge between noisy and clean text. Such a corpus is particularly relevant for this study, as it exposes models to realistic spelling variations while also grounding them in corrected linguistic forms. We present the statistics of the training corpus in Table 1.

3 Experiment Setup

3.1 Dataset Preparation

To systematically evaluate the impact of OOV, we construct two versions of the OLID test set:

- **Raw:** The original dataset with natural noise and spelling variation.
- **Processed:** A normalized version obtained using a spell-checking pipeline (pyspellchecker³).

Additionally, tokens prefixed with @ and # are treated as atomic units to preserve semantic integrity. Moreover, we introduce the metric of Unique Word (UW) percentage to quantify the OOV distribution. Specifically, UW represents the proportion of tokens in the test set that do not appear in the training corpus. This allows for a more fine-grained analysis of model behavior under varying levels of lexical novelty.

3.2 Base Model Training

To ensure comparability, both models share identical architectural configurations: a block size of 64, 4 layers, 4 attention heads, and an embedding dimension of 128. Dropout is disabled to isolate the effects of the modeling paradigm. Due to computational constraints, hyperparameter tuning is not extensively explored. The MLM is trained on CPU, while the DLM is trained on an NVIDIA A100 GPU for approximately 2.5 hours. For the DLM, the sigma frequency dimension is set to 64, following prior work.

3.3 Fine-Tuning and Optimization

Both models are fine-tuned end-to-end on the OLID training set. To address class imbalance, we employ Focal Loss [RD⁺17], which dynamically down-weights easy examples and focuses learning on harder instances. Moreover, the training is conducted with a learning rate of 2×10^{-5} and weight decay of 0.3. Early stopping with a patience of 1 epoch is applied to prevent overfitting. All experiments are conducted with a fixed random seed of 42 to ensure reproducibility.

3.4 Evaluation Metrics

Model performance is evaluated using Precision, Recall, and F1-score. These metrics provide complementary insights:

- **Precision:** Measures the correctness of positive predictions, indicating robustness to noisy or misleading tokens.
- **Recall:** Measures the ability to capture all relevant instances, reflecting sensitivity to diverse linguistic patterns.
- **F1-score:** Provides a harmonic balance between precision and recall.

In the context of OOV, higher precision suggests better contextual disambiguation of noisy tokens, while higher recall indicates improved generalization across unseen or rare word forms.

Category	OFF								NOT								AVG.							
	UW		P		R		F1		UW		P		R		F1		UW		P		R		F1	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
MLM	95.94	95.47	23.68	25.28	23.88	25.10	20.84	23.80	96.40	95.40	55.25	52.65	51.65	51.13	48.68	48.88	96.28	95.68	51.65	50.83	50.49	52.06	46.65	48.10
DLM			27.24	34.01	25.43	25.99	23.42	23.84			62.17	64.54	51.75	52.31	46.42	47.45			59.62	52.25	51.51	65.40	46.56	47.57

Table 2: Unique Word Percentage (UW), Precision (P), Recall (R), and F1 score for each model per class label. The AVG. column indicates the models overall performance.

4 Results & Discussion

Table 2 summarizes the experimental results. A preliminary observation reveals that the raw dataset exhibits a consistently higher UW percentage across both classes, confirming its suitability as an OOV-intensive evaluation setting. Overall, the MLM achieves higher F1-scores across both raw and processed datasets. This suggests that token-level contextual modeling remains highly effective for the classification objective. However, a deeper analysis reveals that the DLM demonstrates notable improvements in both precision and recall, particularly in high-UW settings.

Specifically, the DLM achieves approximately 4% and 7% improvements in precision on datasets with UW levels approaching 96%. This indicates that the diffusion-based approach is better at handling noisy or misspelled inputs, likely due to its character-level modeling and denoising objective. By explicitly learning the transformation between corrupted and clean sequences, the DLM appears to develop a more robust understanding of orthographic variation. In terms of recall, the improvements suggest that the DLM is also more effective at capturing diverse expressions of offensive content, which are often characterized by creative or non-standard spellings. This aligns with the hypothesis that character-level diffusion can enhance generalization in OOV-rich environments. However, the relatively lower F1-score of the DLM indicates a trade-off between precision and overall classification balance. One possible explanation is that while the DLM is better at modeling local variations, it may lack the higher-level semantic abstraction achieved by token-based MLMs. This highlights an important direction for future work: integrating diffusion-based representations with higher-level contextual modeling.

5 Limitation

This study has several limitations. First, both models are trained on a relatively small corpus (approximately 1MB), which may limit their ability to learn rich linguistic patterns. Second, hyperparameter tuning is not extensively explored, which may impact the relative performance of the models. Third, computational constraints result in differences in training hardware (CPU vs GPU), which may introduce subtle biases. Finally, the evaluation is restricted to a single task (OLID), and broader validation across multiple datasets and domains is necessary to generalize the findings.

6 Conclusion

This work presents an initial exploration of character-level diffusion models as a potential solution to the OOV problem in NLP. While MLMs remain strong baselines in terms of overall performance, DLMs demonstrate promising improvements in handling noisy and high-OOV inputs, particularly in terms of precision and recall. The findings suggest that diffusion-based approaches can complement existing token-based methods, offering a new direction for robust language modeling in informal and low-resource settings.

References

- [IKIA21] Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, 2021.
- [KRI25] Roman Kovalchuk, Mariana Romanyshyn, and Petro Ivaniuk. Introducing omnigec: A silver multilingual dataset for grammatical error correction. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 162–178, 2025.
- [Kum25] Ashwani Kumar. The annotated discrete diffusion models. <https://github.com/ash80/diffusion-gpt>, 2025.
- [LME23] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

³<https://pypi.org/project/pyspellchecker/>

- [RD⁺17] T-YLPG Ross, GKHP Dollár, et al. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, volume 488, pages 2980–2988, 2017.
- [ZMN⁺19] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, 2019.