Bridging Distribution Gap via Semantic Rewriting with LLMs to Enhance OOD Robustness

Manas Madine
Department of Computer Science
University of Massachusetts, Amherst
mmadine@umass.edu

Abstract

This paper investigates the robustness of Large Language Models (LLMs) against Out-Of-Distribution (OOD) data within the context of sentiment analysis. Traditional fine-tuning approaches often fail to generalize effectively across different data distributions, limiting the practical deployment of LLMs in dynamic real-world scenarios. To address this challenge, we introduce a novel method called "Semantic Rewriting," which leverages the inherent flexibility of LLMs to align both in-distribution (ID) and OOD data with the LLMs distributions. By semantically transforming sentences to minimize linguistic discrepancies, our approach helps to standardize features across datasets, thus enhancing model robustness. We conduct extensive experiments with several benchmark datasets and LLMs to validate the efficacy of our method. The results demonstrate that Semantic Rewriting significantly improves the performance of models on OOD tasks, outperforming traditional methods in both robustness and generalization capabilities. Our findings suggest that Semantic Rewriting is a promising technique for developing more reliable and versatile NLP systems capable of performing robustly across diverse operational environments.

1 Introduction

In the dynamic field of natural language processing (NLP), Large Language Models (LLMs) have shown exceptional capabilities across a spectrum of applications. Nevertheless, these models frequently encounter challenges with Out-Of-Distribution (OOD) data, which can significantly hinder their effectiveness in varied real-world environments Uppaal et al. (2023); Dai et al. (2023). Conventional methods such as fine-tuning on in-distribution (ID) data often fail to provide robustness against the distribution shifts commonly seen in practical deployments Houlsby et al. (2019).

This paper tackles the critical challenge of bridging the distribution gap between ID and OOD data, essential for the robust deployment of LLMs. Despite considerable advancements in model architectures and training techniques, the issue of distribution shift remains a significant barrier in deploying LLMs across diverse settings Yuan et al. (2024).

Research Questions: This research stems from the research question: whether there exists any projection of ID and OOD data where the distribution gap is minimized, or alternatively, if there exists a global distribution from which we can sample both ID and OOD data.

Contributions: We introduce a novel method called *Semantic Rewriting*, which utilizes the flexibility of LLMs to align their outputs more closely with its own distribution while ensuring semantic equivalence to the original sentences. This approach involves semantically transforming sentences to standardize linguistic properties across ID and OOD datasets, thereby minimizing distributional discrepancies.

We hypothesize that:

- **H1:** A global distribution can bridge the distribution gap between ID and OOD data.
- **H2:** Reducing the distribution shift between ID and OOD data will enhance OOD robustness.

We employ a strategy where both ID and OOD datasets are rewritten through a LLM to standardize their stylistic and semantic features. This process not only promotes homogeneity across datasets but also enables the fine-tuned models on this transformed data to achieve markedly improved performance on OOD tasks. We also use the original ID and fine-tune another instance of RoBerta Liu et al. (2019) which acts as one of our baselines.

Our extensive experiments across benchmark dataset and LLMs validate our approach. The results affirm that semantic rewriting significantly bolsters model robustness against distribution shifts

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 352–362

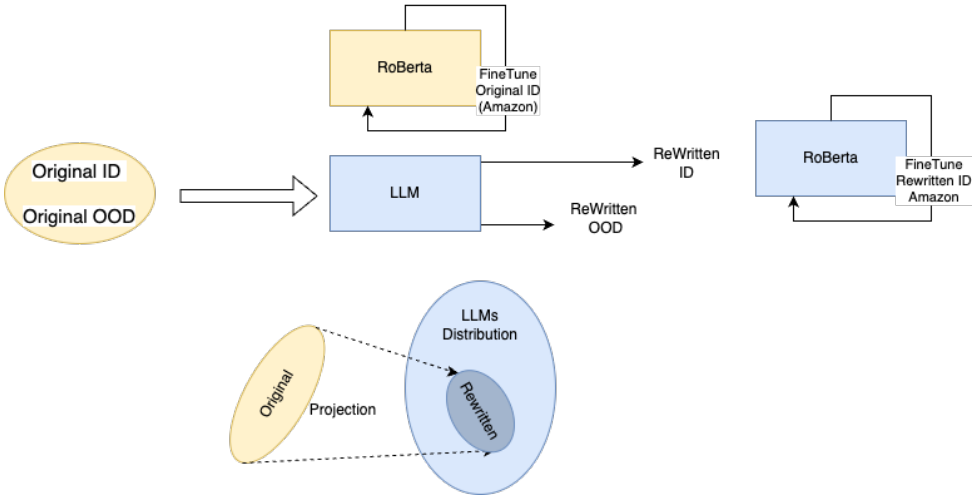

Figure 1: Workflow of Semantic Rewriting for OOD Robustness. The process begins with original ID data is used to fine-tune a RoBERTa model for the baseline, the original in-distribution (ID) and out-of-distribution (OOD) data, which are then transformed by a Large Language Model (LLM) into rewritten forms that align with the LLM's distribution. The rewritten ID data is then used to fine-tune a RoBERTa model, which is then evaluated on the rewritten OOD data. This diagram illustrates the integration of semantic rewriting into the training pipeline to enhance model robustness against distribution shifts.

and, in some instances, enhances performance on ID tasks.

Our work contributes to the broader objective of developing NLP systems that are both robust and versatile, capable of reliably operating across various domains.

2 Related Work

This section reviews relevant literature on the robustness of large language models (LLMs) to out-of-distribution (OOD) scenarios in natural language processing (NLP). Our research is informed by various studies aiming to improve OOD generalization through innovative methods that manipulate data and model interactions.

Bench-marking and Optimization Approaches The "BOSS" benchmark suite introduced by bosspaper is fundamental to our evaluation strategy. It assesses OOD robustness by measuring performance variations across diverse datasets, providing a structured approach to test our semantic rewriting method.

Prompt Optimization and Rewriting The Generalized Prompt Optimization (GPO) framework proposed by Li et al. (2023) utilizes unlabeled target data within prompt optimization to enhance LLM performance on target groups. This concept parallels our semantic rewriting technique where we modify OOD data stylistically to mirror ID data attributes, aiding in bridging the distributional gap.

LLM-based Data Augmentation In alignment with the test-time augmentation strategy of O'Brien et al. (2024), which employs LLMs to generate diverse text augmentations for robustness, our approach uses semantic rewriting to standardize text properties across distributions. This methodology leverages the inherent flexibility of LLMs, suggesting that modifying input text can significantly impact model generalization.

Variational Approaches and Fine-Tuning Zhan et al. (2024) introduces a variational inference framework optimizing the joint distribution of data, contrasting with traditional methods that maximize conditional probabilities. Although different in application, this perspective supports our hypothesis that addressing how data is represented (through rewriting) can mitigate bias introduced by model assumptions. Additionally, Uppaal et al. (2023) questions the necessity of fine-tuning for OOD detection, positing that pre-trained models may already be equipped to handle OOD data effectively, a notion that challenges and inspires our methodology to enhance inherent model capabilities without extensive retraining.

In-Context Learning and Alignment The use of in-context learning (ICL) for style alignment in LLMs, as explored by Lin et al. (2023), directly supports our use of semantic rewriting. Their find-

ings suggest that careful prompt design and example selection can align model output closely with desired outcomes, similar to how we guide LLMs to produce semantically aligned texts.

Comprehensive Approaches Our work builds on the broad analysis by Houlsby et al. (2019), who examine the relationship between performance on ID and OOD datasets through fine-tuning. We extend this by integrating in-context learning techniques, such as prompt engineering and rewriting, to test their efficacy in OOD scenarios without extensive model modifications.
While substantial research focuses on enhancing OOD robustness, few have systematically addressed the use of semantic transformations for this purpose. Our study aims to fill this gap by demonstrating how semantic rewriting, inspired by existing methods, can significantly improve LLMs' OOD robustness. This novel contribution aims to shift the paradigm from model-centric to data-centric approaches in improving OOD generalization.

3 Our Dataset

To evaluate the generalization capabilities of both traditional and modern Language Models (LMs) to Out-Of-Distribution (OOD) data, we focused on sentiment analysis as the primary NLP task. Our study utilizes datasets as specified in the BOSS Benchmark paper, which provides a framework for assessing model performance across different domains Yuan et al. (2024).

We employ one in-distribution (ID) dataset and three OOD datasets, each chosen for their diverse sources and sentiment labeling schemes to comprehensively test the models under varied linguistic contexts. The datasets include:

- **Amazon Reviews (ID)**: This dataset includes reviews across 29 different product categories from Amazon, annotated into three classes—positive, neutral, and negative McAuley and Leskovec (2013).
- **SST-5 (OOD)**: Comprising sentence-level movie reviews from the Rotten Tomatoes website, labeled into the same three sentiment categories Socher et al. (2013).
- **SemEval (OOD)**: A dataset of tweets formatted for sentiment analysis, also segmented into three classes Nakov et al. (2019).

Dataset	Classes	Training	Test
Amazon	3	30,000	38,905
SST-5	3	4,004	1,067
SemEval	3	6,000	20,622
DynaSent	3	93,553	4,320

Table 1: Details of the original datasets for sentiment analysis. The Amazon dataset serves as the in-distribution dataset while SST-5, SemEval, and DynaSent are utilized as out-of-distribution datasets, as per the BOSS Benchmark Yuan et al. (2024).

- **DynaSent (OOD)**: This dataset consists of sentences identified as particularly challenging for sentiment analysis, created using a novel human-and-model-in-the-loop annotation method Potts et al. (2020).

The distribution and structure of these datasets are detailed in Table 1. This selection is instrumental in investigating how well models can adapt when trained on ID data and then tested on data sampled from different, unknown distributions.

For practical purposes and due to computational constraints, we opted to sub-sample the original datasets for our experiments. This process ensured that each class was equally represented, maintaining a balance of sentiment labels across a smaller test dataset. The details of this sub-sampling are presented in Table 2.

Dataset	**Pos**	**Neg**	**Neutral**	**Total**
Amazon	950	950	950	2850
DynaSent	950	950	950	2850
SemEval	950	950	950	2850
SST-5	305	305	305	915

Table 2: Distribution of sentiment labels for the sub-sampled evaluation datasets, ensuring balanced classes across the datasets. The sub-sampling was conducted to facilitate efficient computation while retaining the variability inherent in the original datasets.

4 Theoretical Analysis

In this approach, we are using a Large Language Model (LLM) to rewrite sentences in both in-distribution (ID) and out-of-distribution (OOD) datasets, hypothesizing that this rewriting process will bridge the distribution gap between ID and OOD. Here's a mathematical description and theoretical analysis of why fitting a Gaussian Mixture Model (GMM) for original sentence embedding

would result in K clusters where K is the number of datasets, and rewritten sentence embedding would result in identifying a single cluster.

4.1 Embeddings of Original and Rewritten Sentences

Original Embeddings: Let $\mathbf{X}_{\text{ID}}$ and $\mathbf{X}_{\text{OOD}}$ be the sets of original embeddings from the ID and OOD datasets respectively. Each dataset has its own distribution, leading to K different distributions if there are K datasets.

Rewritten Embeddings: Let $\mathbf{X}'_{\text{ID}}$ and $\mathbf{X}'_{\text{OOD}}$ be the sets of embeddings of the rewritten sentences. We assume these embeddings are produced from the LLM's distribution, denoted as $\mathcal{N}(\mu_{\text{LLM}}, \Sigma_{\text{LLM}})$.

4.2 GMM with Multiple Components

When fitting a GMM with K components, we are essentially assuming the data could be drawn from K different Gaussian distributions. The parameters of each Gaussian component in the GMM are $\theta_k = (\pi_k, \mu_k, \Sigma_k)$, where π_k is the mixing coefficient, μ_k is the mean, and Σ_k is the covariance matrix of the k-th component.

Expectation-Maximization (EM) Algorithm

The EM algorithm iterates between two steps:

E-step: Calculate the responsibility $\gamma(z_{i,k})$ that the i-th data point $\mathbf{x}_i$ belongs to the k-th component.

$$\gamma(z_{i,k}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i|\mu_j, \Sigma_j)}$$

M-step: Update the parameters π_k, μ_k, and Σ_k based on the current responsibilities.

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma(z_{i,k}) \mathbf{x}_i}{\sum_{i=1}^{N} \gamma(z_{i,k})}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma(z_{i,k})(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma(z_{i,k})}$$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma(z_{i,k})$$

4.3 Analysis for Original Sentence Embeddings

Fitting a GMM would result in K components where each component captures one of the K datasets' distributions. As the original embeddings $\mathbf{X}_{\text{ID}}$ and $\mathbf{X}_{\text{OOD}}$ come from K different Gaussian distributions, this leads to multiple clusters.

Convergence to Multiple Clusters: - **E-step**: The responsibilities $\gamma(z_{i,k})$ will reflect the membership of data points to different Gaussian components based on their respective distributions. - **M-step**: The parameters μ_k and Σ_k of each component will converge to the mean and covariance of the respective distributions of the original datasets. The mixing coefficients π_k will reflect the proportion of points belonging to each dataset.

Given the original embeddings:

$$\mathbf{x}_i \sim \mathcal{N}(\mu_{\text{ID}_k}, \Sigma_{\text{ID}_k}) \quad \text{for} \quad k = 1, \ldots, K$$

Fitting a GMM with K components will result in:

$$\mu_k \approx \mu_{\text{ID}_k}, \quad \Sigma_k \approx \Sigma_{\text{ID}_k}, \quad \pi_k \approx \frac{\text{size of } \mathbf{X}_{\text{ID}_k}}{N}$$

for each of the K components, where $\mathbf{X}_{\text{ID}_k}$ represents embeddings from the k-th dataset.

4.4 Analysis for Rewritten Sentence Embeddings

If all embeddings $\mathbf{X}'_{\text{ID}} \cup \mathbf{X}'_{\text{OOD}}$ are produced by a single Gaussian distribution $\mathcal{N}(\mu_{\text{LLM}}, \Sigma_{\text{LLM}})$, the responsibilities $\gamma(z_{i,k})$ for the component that best fits $\mathcal{N}(\mu_{\text{LLM}}, \Sigma_{\text{LLM}})$ will be near 1, and for other components, they will be near 0.

Convergence to One Cluster: - **E-step**: Responsibilities $\gamma(z_{i,k})$ will indicate that all points $\mathbf{x}'_i$ mostly belong to one Gaussian component. - **M-step**: The parameters μ_k and Σ_k of this dominant component will converge to μ_{LLM} and Σ_{LLM}. The mixing coefficient π_k will converge to 1 for this component and 0 for others.

Given the rewritten embeddings:

$$\mathbf{x}'_i \sim \mathcal{N}(\mu_{\text{LLM}}, \Sigma_{\text{LLM}})$$

When fitting a GMM with K components to $\mathbf{X}'_{\text{ID}} \cup \mathbf{X}'_{\text{OOD}}$, the maximum likelihood estimate will find that:

$$\mu_k \approx \mu_{\text{LLM}}, \quad \Sigma_k \approx \Sigma_{\text{LLM}}, \quad \pi_k \approx 1$$

for one component, and the responsibilities for other components will be negligible.

4.5 Analysis Summary

By rewriting the sentences using an LLM, the embeddings of both ID and OOD datasets become

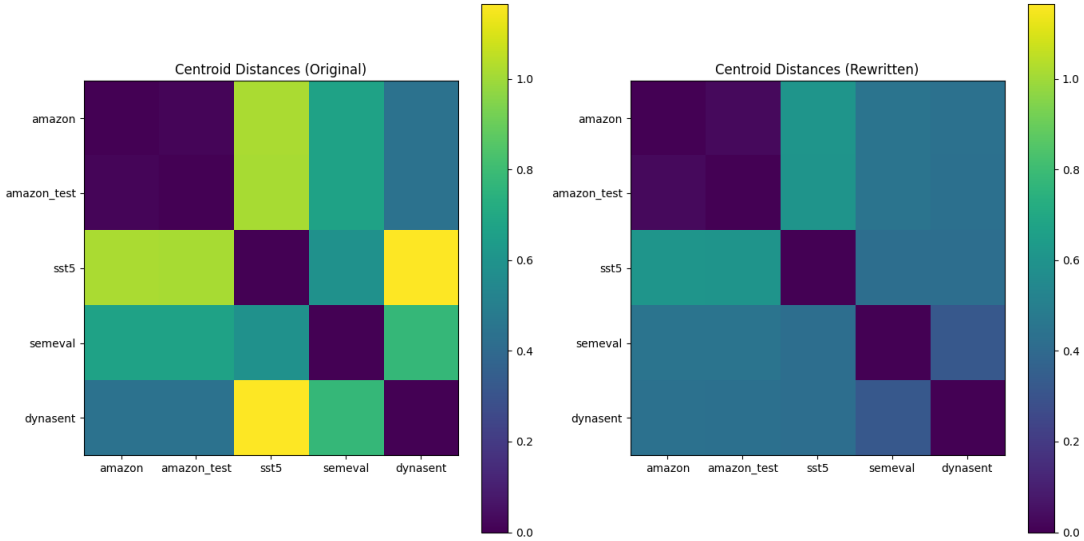

Figure 2: Comparison of centroid distances between original and rewritten embeddings across various datasets.

samples from the same underlying Gaussian distribution $\mathcal{N}(\mu_{\text{LLM}}, \Sigma_{\text{LLM}})$. Consequently, when fitting a GMM, the algorithm identifies a single cluster that represents the LLM's distribution. This theoretical foundation supports empirical observations that a RoBERTa model fine-tuned on these rewritten sentences performs better on OOD data(details in further sections). Conversely, the original sentence embeddings will result in multiple clusters, reflecting the diverse distributions of the original datasets.

5 Experimental Validation

In this section, we validate our theoretical framework by presenting empirical results obtained from the application of our sentence rewriting strategy using a Large Language Model (LLM). We analyze the impact of rewriting on the distribution of sentence embeddings from various datasets.

5.1 Centroid Distance Analysis

The centroid distances between different datasets before and after the rewriting process were computed to quantify the distribution shifts. As depicted in Figure 2, the centroid distances among datasets such as Amazon, SST-5, SemEval, and DynaSent are reduced significantly after the rewriting process. This indicates a closer alignment of distributions, supporting our hypothesis that rewriting can effectively minimize distributional discrepancies. Refer to Figure 2 for details.

5.2 UMAP Visualization of Embeddings

To visualize the effect of rewriting on the embedding space, we utilized UMAP to reduce the dimensionality of embeddings to two dimensions. The UMAP plots, shown in Figure 3, clearly demonstrate a more cohesive and overlapping distribution of embeddings after rewriting. The original embeddings exhibit distinct clusters corresponding to different datasets, whereas the rewritten embeddings tend to form a single, unified cluster, further validating the effectiveness of our approach in reducing distribution shifts.

5.3 Cluster Validity Analysis

We fitted a Gaussian Mixture Model (GMM) to both original and semantically rewritten embeddings and used the Akaike Information Criterion (AIC) and Silhouette scores to determine the optimal number of clusters. Lower AIC scores indicate a better-fitting model by balancing fit and complexity, while higher Silhouette scores (ranging from -1 to 1) indicate well-separated clusters. Figure 4 shows that rewritten embeddings require fewer components, with a single cluster being the most fitting, supporting our hypothesis that semantic rewriting aligns data distributions.

Our experiments confirm that semantic rewriting with LLMs significantly harmonizes sentence embeddings across different datasets. This is demonstrated by reduced centroid distances, cohesive UMAP visualizations, and simplified GMM clustering, highlighting the potential of this approach to enhance model generalization.

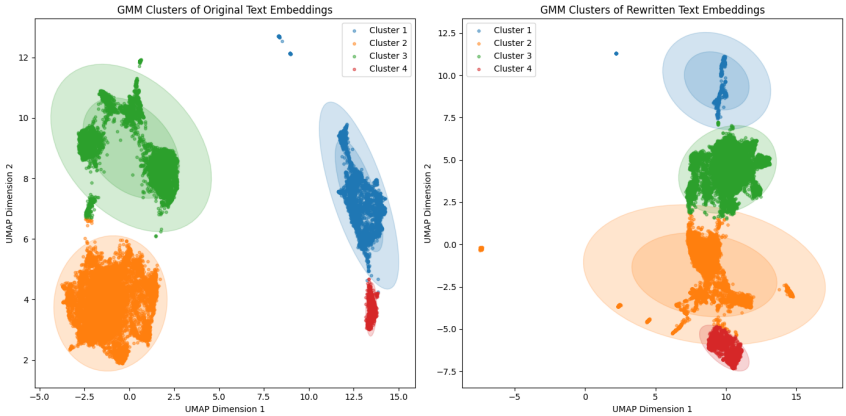

Figure 3: UMAP visualizations of original and rewritten text embeddings showing the distributional shift and clustering behavior.

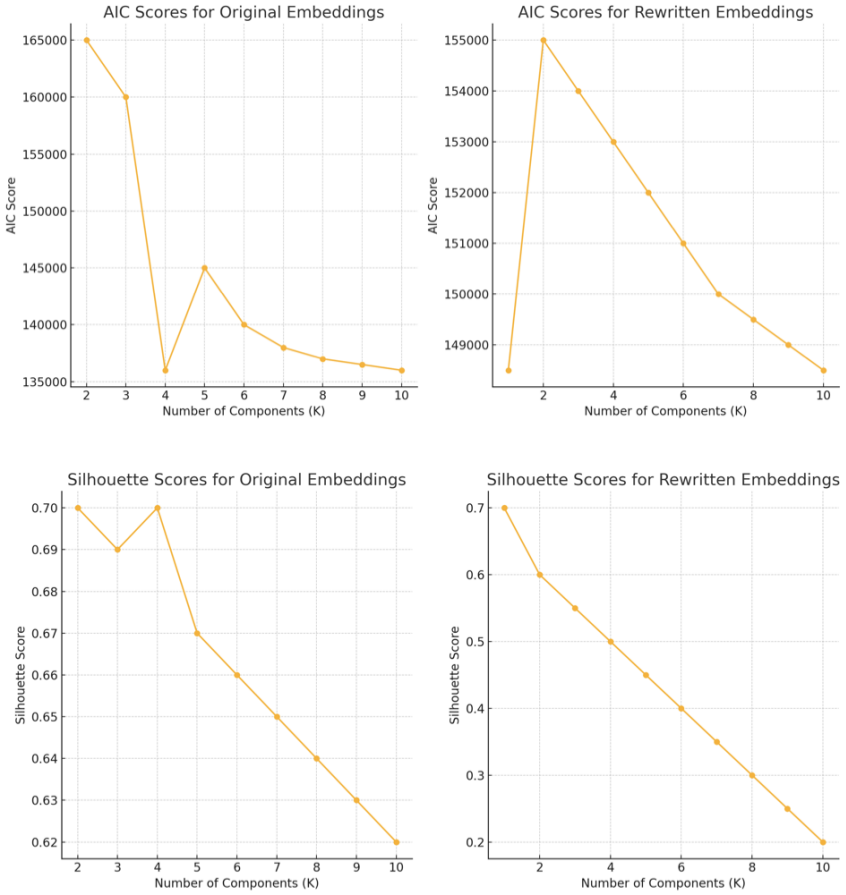

Figure 4: AIC and Silhouette scores for original and rewritten embeddings, supporting a reduced number of clusters post-rewriting.

6 Baselines

In this study, we evaluate the generalization capabilities of various Language Models (LMs) to Out-Of-Distribution (OOD) datasets using different methods of prompting and rewriting. Our baselines include traditional fine-tuning and more recent approaches like zero-shot prompting, and a novel method we introduce: semantic rewriting. We compare these methods across multiple datasets, including Amazon as the In-Distribution (ID) dataset and DynaSent, SST-5, and SemEval as OOD datasets.

6.1 Traditional Fine-Tuning

We first assess the performance of RoBERTa, a robustly optimized BERT pretraining approach,

fine-tuned on the Amazon dataset. This serves as our conventional baseline. RoBERTa demonstrates strong performance on the ID dataset (Amazon), but shows less and varied performance on OOD datasets, highlighting challenges in handling distribution shifts.

6.2 Zero-Shot Prompting

Expanding our investigation into the efficacy of zero-shot capabilities, we utilize variants of the LLaMA model. We used Llama 3 8B and its 4 bit quantized version and also Llama 2 70B models. The LLaMA models, known for their few-shot learning prowess, are tested in a zero-shot setup where they directly predict sentiment without fine-tuning.

6.3 In-Context Rewriting

Based on the O'Brien et al. (2024) , this method involves using in-context learning to rewrite OOD and ID-test data samples to resemble ID samples, leveraging samples from ID as a template. The LLM is prompted to generate text that aligns with the ID data, after which it performs zero-shot classification on these rewritten texts. This approach explores how well LLMs can adapt their output to match the distribution of ID data when generating OOD samples.

7 Method

In order to improve performance over baseline approaches, we propose semantic-rewriting followed by fine-tuning RoBerta Liu et al. (2019) with the ID rewritten data and evaluate the model's performance on OOD rewritten data.

Semantic-Rewritting Given a LLM (Large language model), L1, we feed inputs for the in distribution dataset and prompt it in a zero-shot manner to semantically rewrite the ID sentences (Amazon). This step involves mapping the rewritten sentence within the LLMs distribution. Similarly we do the semantic rewriting of the OOD datasets(SemEval, Dynasent, SST-5) and also Amazon test using the same LLM to map the responses to the LLMs distribution there by bridging the distributing gap of the ID and OOD datasets as now they come from the same distribution as that of the LLM.

7.1 Implementation Details

Data Pre-processing and Fine-Tuning: Datasets were obtained from the BOSS Benchmark Yuan et al. (2024) and pre-processed to ensure uniformity across training and testing samples. Fine-tuning was performed using the RoBERTa model on the Amazon Reviews dataset, which included 9,000 training samples sub-sampled from the BOSS Benchmark (see Table 1 for details), with a 10% validation split. The models were trained for 5 epochs with a batch size of 32 and a learning rate of 2e-5, using the AdamW optimizer with a linear scheduler. For our test data, we sub-sampled from the original test sets as shown in Table 2. In our semantic rewriting method, we rewrote the same 9,000 training samples and fine-tuned a RoBERTa model. For testing, we used the semantically rewritten Amazon, SST-5, SemEval, and DynaSent test sets.

Prompting Techniques: We employed different prompting techniques using Llama-2-70B, Llama-3-8B, and Llama-3-8B-4Bit models. Zero-shot and In-context-rewriting were tested, along with our novel semantic rewriting strategy. For inference, we utilized the Together AI API AI (2024) for Llama models and conducted zero-shot classification and rewriting using prompts designed to enhance sentiment analysis accuracy.

Computational Resources: Fine-tuning and evaluations were performed on L4 , T4 and A100 GPUs from colab pro, with the A100 40GB model reducing the epoch duration to approximately 1 hour. Monitoring and logging of model training were facilitated by the WandB platform Wan (2024).

Code Availability: The code used for all experiments has been made publicly available for reproducibility and further research at our code (2024).

8 Results

This section presents the findings from our experimental validation, comparing the performance of various models and methods on both in-distribution (ID) and out-of-distribution (OOD) datasets. The methods evaluated include traditional fine-tuning, zero-shot learning with different LLaMA models, and our novel semantic rewriting approach.

Method	Model	Amazon*	DynaSent	SST-5	SemEval
Original	RoBERTa	82.64%	63.79%	58.45%	40.21%
zero-shot	Llama 3 (8 B 4 bit)	72.44%	67.88%	69.82%	57.22%
	Llama 3 (8 B)	74.22%	66.77%	70.96%	62.00%
	Llama 2 (70 B)	78.55%	63.55%	72.06%	63.66%
In-Context Rewriting	Llama 3 (8 B 4 bit)	46.00%	30.67%	38.58%	34.78%
	Llama 3 (8 B)	50.00%	38.67%	39.65%	38.00%
	Llama 2 (70 B)	64.47%	54.44%	60.60%	52.89%
Semantic Rewritting	RoBERTa	**84.91%**	**76.99%**	**74.22%**	**67.22%**

Table 3: Performance of prompting techniques on different datasets. The models perform well on OOD tasks compared to simple fine-tuning.
*ID test datasets

8.1 Traditional Fine-Tuning Performance

Our baseline method using the RoBERTa model fine-tuned on the Amazon dataset (ID) achieved an accuracy of 82.64%. However, its performance on the OOD datasets was less robust, scoring 63.79% on DynaSent, 58.45% on SST-5, and 40.21% on SemEval. These results highlight the limitations of traditional fine-tuning methods in handling distribution shifts effectively.

8.2 Zero-Shot Learning Performance

The zero-shot learning method was tested using various configurations of the LLaMA model. The results are as follows:

- LLaMA 3 (8B 4 bit) achieved 72.44% on Amazon and showed moderate improvement on OOD datasets with 67.88% on DynaSent, 69.82% on SST-5, and 57.22% on SemEval.
- LLaMA 3 (8B) scored slightly higher with 74.22% on Amazon and comparable results on OOD datasets.
- The larger LLaMA 2 (70B) model outperformed the smaller versions on Amazon with 78.55% and demonstrated the best OOD performance, particularly on SST-5 with 72.06% and SemEval with 63.66%.

These findings underscore the potential of zero-shot learning with large-scale models to adapt better to OOD scenarios without the need for extensive retraining.

8.3 Performance of In-Context Rewriting

The in-context rewriting approach, inspired by the O'Brien et al. (2024) paper, leveraged the ID dataset to generate rewritten OOD samples that mimic the ID distribution. This approach did not fare well compared to zero-shot learning, indicating that while the model could generate stylistically similar outputs, the semantic content adaptation was less effective for OOD generalization.

8.4 Semantic Rewriting Performance

Our semantic rewriting method, which involved retraining RoBERTa on semantically rewritten ID and OOD datasets, showed significant improvements:

- On Amazon, it achieved the highest accuracy of 84.91%.
- It dramatically improved OOD robustness with 76.99% on DynaSent, 74.22% on SST-5, and 67.22% on SemEval.

These results validate our hypothesis that semantic rewriting can bridge the distribution gap between ID and OOD data, enhancing the model's overall robustness and performance across varied datasets.

The experiments confirm that while traditional methods and zero-shot learning provide foundational capabilities, advanced techniques like semantic rewriting offer substantial improvements in model robustness and OOD generalization. This approach not only aligns ID and OOD distributions more closely but also preserves and even enhances performance on ID tasks, establishing a new benchmark for future research in OOD robustness in NLP.

9 Conclusion

Our extensive experiments with several benchmark datasets and LLMs demonstrated that Semantic Rewriting significantly improves the performance

of models on OOD tasks, outperforming traditional methods in both robustness and generalization capabilities. The results indicated a substantial reduction in centroid distances, more cohesive UMAP visualizations, and simplified GMM clustering, highlighting the potential of this approach to enhance model generalization.

10 Limitations

Generality of Approach: Although our approach shows significant improvements in the context of sentiment analysis, its generalizability to other NLP tasks remains to be explored. Different tasks may require tailored rewriting strategies.

Model Diversity: Our experiments primarily utilized RoBERTa and LLaMA models. To validate the broader applicability of Semantic Rewriting, additional testing on a diverse range of models, such as BERT, GPT, T5, and other transformer-based architectures, is necessary. This would help ascertain the method's effectiveness across different model architectures and configurations.

References

2024. Wandb. https://wandb.ai/home. Accessed: 2024-05-17.

Together AI. 2024. Api access. https://docs.together.ai/docs/quickstart. Accessed: 2024-05-17.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. 2023. Exploring large language models for multi-modal out-of-distribution detection. *arXiv preprint arXiv:2310.08027*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. 2023. Robust prompt optimization for large language models against distribution shifts. Association for Computational Linguistics.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2019. Semeval-2016 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.01973*.

Kyle O'Brien, Nathan Ng, Isha Puri, Jorge Mendez, Hamid Palangi, Yoon Kim, Marzyeh Ghassemi, and Thomas Hartvigsen. 2024. Improving black-box robustness with in-context rewriting. *arXiv preprint arXiv:2402.08225*.

our code. 2024. Robust-llm-to-ood. https://github.com/manas1999/Robust_LLMS_TO_OOD.git. Accessed: 2024-05-17.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. Dynasent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Rheeya Uppaal, Junjie Hu, and Yixuan Li. 2023. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. *arXiv preprint arXiv:2305.13282*.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2024. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36.

Li-Ming Zhan, Bo Liu, and Xiao-Ming Wu. 2024. Vi-ood: A unified representation learning framework for textual out-of-distribution detection. *arXiv preprint arXiv:2404.06217*.

A Appendix A:

A.1 Zero-Shot Prompt

zero-shot Prompt

Instructions
For sentiment analysis: Your task is to perform a sentiment analysis on a given input text and provide a single word indicating whether the sentiment is positive, negative, or neutral. The input text may contain any language or style of writing. Please ensure that your analysis takes into account the overall tone and context of the text. Your response should be concise and clear, providing a single word that accurately reflects the sentiment of the input text. If there are multiple sentiments present in the text, please choose the one that best represents the overall feeling conveyed by the author. Please note that your analysis should take into account all relevant factors, such as tone, language use, and content. Your response should also be flexible enough to allow for various types of input texts.

We used the same prompt as Li et al. (2023) paper.

A.2 In context Rewriting Prompt

Rewriting

Instructions

The assistant is to paraphrase the input text as if it was one of the examples. Change the details of the text if necessary.
Style Examples
$< style\_transfer\_exemplars >$

We used the same prompt as O'Brien et al. (2024) paper, We used 7 samples from the ID in the prompt as In-context examples.

A.3 Semantic Rewritting

LLaMA Prompt template (Unsloth)

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request
Instructions
{Please rewrite the sentence to standardize its style, tone, and format.The rewritten sentence should be neutral in tone, concise, and focus on the essential aspects of the sentence only and remove any styles and personal anecdotes. Adjust any colloquial language to a more formal tone. Your goal is to make the sentence indistinguishable in terms of origin, whether it be Amazon, SST-5, or any other dataset.Your rewritten sentence should begin with ¨Rewritten Sentence: ... .¨}

Input Text
{}

Output Text
{}

B Appendix B: Second Appendix

B.1 LLM Inference Parameters

The following hyper-parameters were used for the LLM inference:

- **temperature**: The sampling temperature, set to `0.7`. This parameter controls the randomness of predictions by scaling the logits before applying softmax. Lower values make the model more conservative, while higher values increase randomness.
- **top_p**: The nucleus sampling probability, set to `0.7`. This parameter specifies the cumulative probability threshold for nucleus sampling, where only the smallest set of most probable tokens with probabilities summing up to `top_p` are considered.
- **top_k**: The number of highest probability vocabulary tokens to keep for top-k sampling, set to `50`. This parameter limits the sampling pool to the top-k tokens, reducing the probability mass considered during generation to the top `top_k` tokens.
- **repetition_penalty**: The penalty for repeated sequences, set to `1`. This parameter penalizes repeated tokens in the sequence, encouraging the model to produce more diverse outputs.