Utility-Based Preference Training for Effective Synthetic Text Classification

Anonymous ACL submission

Abstract

We propose a novel approach for generating high-quality synthetic text data for multiclass text classification by leveraging large language models (LLMs) with preference-based fine-tuning. Our method modifies the Direct Preference Optimization (DPO) framework by incorporating a margin-based utility signal that encourages class-discriminative text generation. This margin-based variant, which we call Utility DPO (U-DPO), promotes the generation of synthetic samples with clearer label-specific features.We evaluate our method on two academic document classification benchmarks, Arxiv and WOS-11967, which cover 11 and 33 classes, respectively. Synthetic data generated by a language model trained with U-DPO leads to better classification performance than data generated by a baseline LLM or a model trained with standard DPO. Notably, U-DPO yields consistent improvements in classification accuracy, both when models are trained exclusively on synthetic data and when synthetic data is used to augment limited real data, highlighting the practical value of preference-optimized synthetic datasets. In general, our work demonstrates that incorporating task-specific utility signals into LLM training is a promising direction to generate effective synthetic data for text classification, enabling improved downstream performance without additional human annotation.

1 Introduction

Text classification is a fundamental task in natural language processing, and recent advances in Large Language Models (LLMs) have opened new possibilities to address it (Li et al., 2023; Wang et al., 2024; Kostina et al., 2025). In particular, using LLMs to generate labeled synthetic text has emerged as an attractive approach to supplement or even replace real training data in scenarios with limited annotations (Yoo et al., 2021; Kruschwitz and Schmidhuber, 2024). However, a key challenge remains: ensuring that the generated texts are label-specific and useful for training robust classifiers (Li et al., 2023). Unguided text generation can produce outputs that do not reflect class-specific distinctions, limiting their value for supervised learning (Yamagishi and Nakamura, 2024; Nadas et al., 2025; Gan and Liu, 2025). In this paper, we address the above challenge by introducing a preference-based framework for synthetic data generation tailored to text classification tasks. We build on Direct Preference Optimization (DPO) (Rafailov et al., 2024), originally proposed to align LLMs with human preferences. Our approach, called Utility DPO (U-DPO), modifies DPO for class-conditional text generation. In U-DPO, the LLM is fine-tuned using preference signals that favor outputs with stronger class relevance: for each class, the model learns to prefer candidate generations that better exhibit the distinctive characteristics of the label. By explicitly optimizing for label consistency and discriminative content, our method produces synthetic examples that are more aligned with downstream classification needs than those from conventional prompting or preference alignment alone. We evaluate the proposed U-DPO approach on two multiclass text classification datasets of research documents, comparing it against baseline synthetic data generation and standard DPO tuning. Our experiments show that U-DPO synthetic data leads to consistently better classification performance than baseline synthetic data. Moreover, when a modest amount of real data is available, augmenting it with U-DPO synthetic samples yields further improvements over using real data alone or with other synthetic data.We also performed an analysis using a margin-based confidence metric to verify that U-DPO indeed produces more label-consistent text. Our analysis confirms that U-DPO samples have significantly higher margin scores on average, indicating a stronger class signal in the generated content.

Our main contributions are as follows.

We extend the DPO preference-based fine-tuning framework to the domain of synthetic data generation for classification, introducing methods that promote class-discriminative text generation.

We propose Utility DPO (U-DPO), which incorporates task-specific utility signals based on classification preferences into LLM fine-tuning, resulting in more informative synthetic training data.

Through extensive experiments, we show that U-DPO improves downstream classifier performance, often narrowing the gap between real and synthetic training data. U-DPO also proves to be effective in hybrid settings, where synthetic and real data are combined.

2 Related Work

2.1 LLMs for Text Classification

Recent studies explore the capabilities of large language models (LLMs) in text classification through zero-shot and few-shot prompting as well as fine-tuning (Wang et al., 2023; Meshkin et al., 2024). These works show that LLMs can often perform surprisingly well in classification tasks without task-specific training data, but their effectiveness varies by task and setting (Bucher and Martini, 2024). LLMs have shown competitive performance in specialized tasks such as scientific edit intent classification (Ruan et al., 2024), but large-scale evaluations report that zero-shot prompting is effective mainly on simple tasks like sentiment analysis, while fine-tuned models remain stronger on more complex classification problems (Vajjala and Shimangaud, 2025). Moreover, a recent multilingual study found that smaller fine-tuned transformers can even surpass few-shot LLMs in accuracy across most categories, suggesting that in-context learning alone is often insufficient for optimal classification performance (Edwards and Camacho-Collados, 2024).

2.2 Prompt-Based Synthetic Data Generation

Prompt-based synthetic data generation has emerged as a promising strategy for training text classifiers (Yoo et al., 2021). Instead of manually collecting or annotating data, researchers prompt LLMs to produce labeled examples, which can be used to augment or even replace human-labeled training sets (Li et al., 2023). Such approaches have shown growing effectiveness for domain adaptation and general-purpose classification (Tan et al., 2024). For instance, recent studies show that LLMs can generate domain-general sentiment datasets (Choi et al., 2024), fully synthetic training corpora without human labels (Peng et al., 2024), and even code-mixed data for multilingual sentiment classification (Zeng, 2024). Despite these successes, important limitations have also been observed. In one health-related classification task, augmenting an unbalanced dataset with GPT-generated samples did not produce performance improvements (Yamagishi and Nakamura, 2024). This suggests that the effectiveness of synthetic data depends on the target data, and that generation strategies must be carefully tailored.

2.3 Preference Optimization for Text Generation

The standard approach, Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2023), optimizes a model to produce outputs preferred by humans and has been widely used to train aligned language models (Stiennon et al., 2022; Ouyang et al., 2022). Some prior work has also explored application of reinforcement learning to structured prediction tasks such as text classification (Chai et al., 2020; Sharma et al., 2025). However, RLHF can be complex and unstable, especially in classification settings, where defining reliable reward functions can be challenging (Kaufmann et al., 2024). To address these issues, Direct Preference Optimization (DPO) avoids reward modeling and learns directly from human preferences (Rafailov et al., 2024). Building on this idea, subsequent work has adapted DPO to more complex alignment tasks. For instance, one line of work extends DPO to multi-turn dialogue settings by introducing a sequential objective tailored for conversational agents (Shi et al., 2025). Another approach improves calibration by aligning model scores with human reward scales (Xiao et al., 2024). Together, these approaches suggest that preference optimization can be effectively adapted to diverse task requirements, offering a practical and scalable framework for alignment.

3 Method

In this section, we present our modified Direct Preference Optimization (DPO) (Rafailov et al., 2024) framework for class-conditional synthetic text generation. Unlike the original DPO formulation, which is designed for preference alignment in instruction tuning, our approach explicitly en-

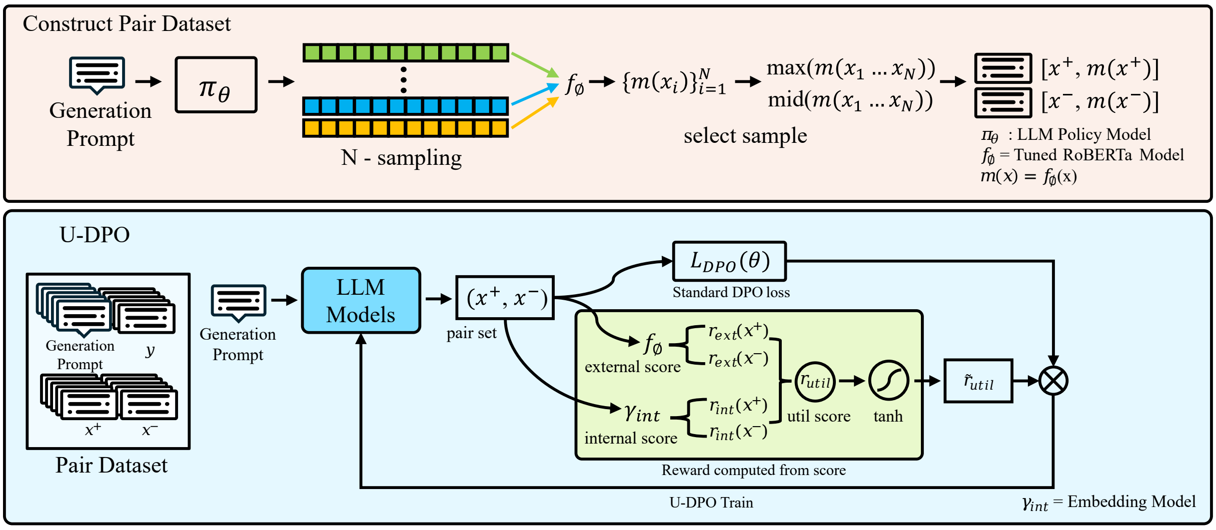

Figure 1: Overview of our modified Direct Preference Optimization (DPO) framework for class-conditional synthetic text generation.

courages class-discriminative generation suitable for multiclass classification tasks. An overview of the entire framework is illustrated in Figure 1.

3.1 Problem Formulation

Let $\mathcal{C} = \{c_1, \dots, c_K\}$ be the set of K class labels. Our goal is to train a language model π_θ that can generate synthetic text x conditioned on a given class label $c \in \mathcal{C}$ such that the generated data are useful for training downstream classifiers. To achieve this, we adopt a preference-based learning setup, where for each class c, a pair of candidate generations (x^+, x^-) is assumed to reflect relative quality under the class semantics, where x^+ denotes the preferred sample and x^- the less preferred one.

3.2 Limitations of Standard DPO

The DPO objective is defined over preference pairs (x^+, x^-), optimizing the model to prefer x^+ over x^- without explicitly considering the underlying task structure. Formally, it minimizes the following loss.

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\theta) = & -\log \sigma\big(\beta \cdot \log \pi_\theta(x^+) \\ & -(1-\beta) \cdot \log \pi_\theta(x^-)\big)\end{aligned} \tag{1}$$

While this formulation effectively aligns model outputs with human preferences in open-ended generation tasks, it presents key limitations when applied to classification. Since DPO training aligns with human preferences without regard to downstream task structure, it can generate outputs that lack label-specific discriminative features. As a result, the model may generate outputs that are linguistically well-formed but lack clear label-specific signals required for classification.

3.3 Data Construction for DPO Training

A key challenge in applying DPO to classification is constructing preference pairs that reflect the relevance of the label. Since human-labeled pairs are costly, we generate preference pairs automatically using a lightweight classifier f_ϕ. Specifically, we first sample n candidate texts for each class using the language model π_θ, and then compute margin scores for each sample using f_ϕ. Based on these scores, we construct preference pairs by selecting high-scoring samples as x^+ and mid-scoring samples as x^-. The margin score for a given generated sample $\tilde{x}$ and target class y is defined as:

$$m(\tilde{x}) = f_\phi(\tilde{x})_y - \max_{j \neq y} f_\phi(\tilde{x})_j \tag{2}$$

where $f_\phi(\tilde{x})_y$ denotes the predicted probability for class y, and the second term represents the highest predicted probability among all other classes.

3.4 Reward Utility Function Design

Although the preference pairs are constructed using margin-based scores to reflect class relevance, relying solely on margin values during training can be risky. Margin scores from the auxiliary classifier may be noisy or biased, and do not capture semantic quality such as fluency or coherence. To address this, we define a reward utility function $r_{\text{util}}(\cdot)$ that combines two complementary signals:

an internal score $r_{\text{int}}(x)$, capturing semantic quality based on embedding similarity, and an external score $r_{\text{ext}}(x, y)$, reflecting label-level confidence via classifier margin. This combination encourages the model to balance linguistic fluency and class relevance, mitigating over-reliance on potentially biased or noisy classifier outputs.

$$r_{\text{ext}}(\tilde{x}, y) = f_\phi(\tilde{x})_y - \max_{j \neq y} f_\phi(\tilde{x})_j \tag{3}$$

$$r_{\text{util}}(\tilde{x}, y) = \lambda \cdot r_{\text{int}}(\tilde{x}) + (1 - \lambda) \cdot r_{\text{ext}}(\tilde{x}, y) \tag{4}$$

However, the definition of a utility-based reward function alone does not guarantee that all preference pairs provide meaningful learning signals. Pairs that are comparably strong or weak often lack meaningful contrast, despite having an assigned preference. To reduce the influence of such cases, we apply a modulation factor based on the r_{util}. Specifically, we adopt the tanh function for its stability and boundedness, enabling smooth scaling from 0 to 1.

$$w(x^+, x^-) = \tanh\left(\left|r_{\text{util}}(x^+, y) - r_{\text{util}}(x^-, y)\right|\right) \tag{5}$$

The final DPO training loss is computed by weighting each pair's contribution according to the modulation factor:

$$\hat{\mathcal{L}}_{\text{DPO}}(\theta) = \mathcal{L}_{\text{DPO}}(\theta) \cdot w(x^+, x^-) \tag{6}$$

4 Experiments

4.1 Experimental Setup

We evaluate our approach on two multiclass scientific classification datasets: *Arxiv* (Clement et al., 2019) and *WOS-11967* (Kowsari et al., 2017), both consisting of scholarly abstracts.

The overall experimental configuration is summarized in Table 1. We use *SciBERT (uncased)* (Beltagy et al., 2019) as the classifier backbone and *MiniLM-L12-H384-uncased* (Wang et al., 2020) as the embedding model for computing semantic utility scores. Prior studies have shown that SciBERT consistently outperforms BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) on scientific NLP benchmarks, highlighting its domain relevance and strong baseline performance. Accordingly, we adopt SciBERT as the backbone for our experiments.

Synthetic training data is generated using three open-source LLMs— LLaMA 3.2 1B, 3B (Grattafiori et al., 2024), and Phi-4-mini (Microsoft et al., 2025)—with class-conditional prompts and utility-based filtering. DPO training is performed using HuggingFace TRL with custom modifications to incorporate utility-aware pair selection. All experiments are conducted on a single NVIDIA A6000 GPU with 48GB of memory.

Synthetic data are generated using a 2-shot prompting strategy, where two randomly selected examples with the same label are used as input to the LLM. For each prompt, we generate $n = 5$ samples to encourage diversity while preserving class consistency. For evaluation, we report accuracy score.

Table 1: Summary of experimental configuration.

Dataset	Arxiv	WOS-11967
category	11	33
train set	28388	9573
test set	2500	2394

Component	**Details**
Classification Model	SciBERT
Embedding Model	MiniLM
LLMs for Generation	LLaMA 3.2 1B, LLaMA 3.2 3B, Phi-4-mini 3.8B
DPO Settings	$\beta = 1.0$, lr = 2e-5, batch size = 2, epochs = 3

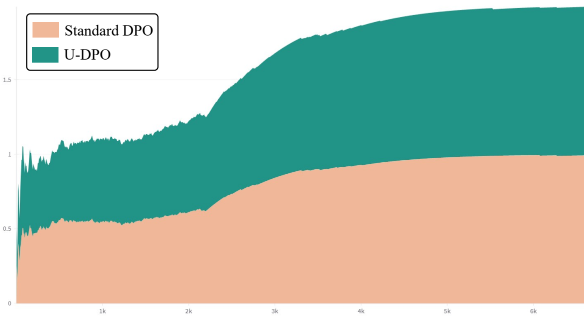

Figure 2: DPO reward accuracy over the course of training for standard DPO and U-DPO.

4.2 Training-time Preference Consistency

To assess how well the model aligns with the preference signal during training, we compute the *DPO reward accuracy*—defined as the percentage of training pairs (x^+, x^-) for which the current model assigns a higher reward (log-probability) to the preferred sample x^+.

As shown in Figure 2, U-DPO maintains consistently higher reward accuracy throughout train-

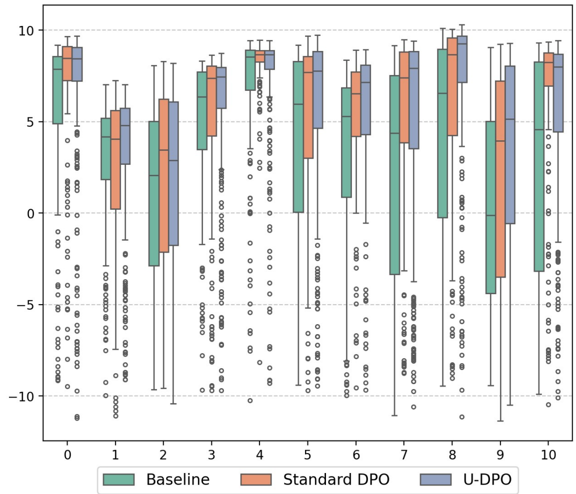

Figure 3: Margin score distributions for synthetic samples generated from identical prompts using the `Phi-4-mini` model on the `Arxiv` dataset. We compare three training regimes: baseline (no preference optimization), standard DPO, and U-DPO.

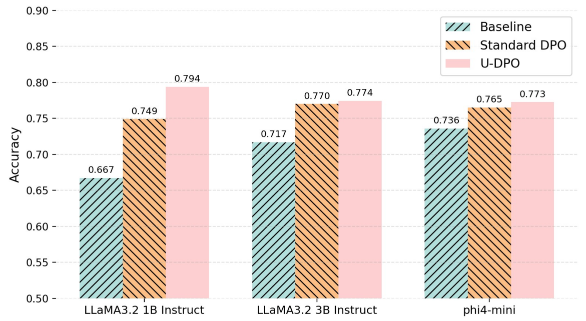

Figure 4: Downstream classification accuracy of models trained exclusively on synthetic data generated under three training regimes: baseline (no preference optimization), standard DPO, and U-DPO. All models are evaluated on the `Arxiv` test set.

ing compared to standard DPO. This indicates that utility-filtered preference pairs are better aligned with the model's learning signal, enabling more efficient and stable optimization. In contrast, standard DPO exhibits greater fluctuations in reward accuracy, likely due to inconsistencies and noise introduced by unfiltered pair selection. The area plot shows that U-DPO achieves consistently higher alignment with the preference signal, indicating more stable and effective training dynamics.

4.3 Margin-Based Quality Assessment

To verify whether our U-DPO training framework enhances the class consistency of generated text, we evaluate the quality of synthetic samples using a lightweight classifier, following the margin score definition introduced in Section 3.4. The margin score quantifies the model's confidence in the correct label over the second-best prediction, with higher values indicating stronger class discriminability.

We analyze this margin score using samples generated by the `Phi-4-mini` model on the `Arxiv` dataset. As shown in Figure 3, both standard DPO and our U-DPO training regimes consistently result in higher average and median margin scores compared to generation without preference optimization, indicating that preference-based training improves label alignment. Notably, U-DPO yields further gains by producing text with higher margin scores in most cases, suggesting that incorporating utility signals during training strengthens the model's ability to generate label-consistent outputs. However, we note that margin scores do not always directly translate into improved downstream classification performance, as they primarily indicate confidence at the sample level rather than overall task-level generalization. Additional results on other model–dataset combinations are provided in Appendix A.

4.4 Classification Performance with Synthetic Data

To directly assess the classification performance of synthetic data, we train models exclusively on generated samples from the `Arxiv` dataset and evaluate their ability to generalize to real-world tasks. This setup allows us to examine whether improvements in sample-level quality lead to better generalization on real-world tasks.

We compare three training regimes for synthetic data generation: (1) generation from the base LLM without preference training, (2) generation using standard DPO, and (3) generation via our proposed **U-DPO** framework. For each setting, a classifier is trained solely on the generated data and evaluated on the original test dataset using **accuracy** as the primary metric. As shown in Figure 4, classifiers trained using U-DPO samples consistently outperform those trained on data from the base LLM and standard DPO. These results demonstrate that utility-based training not only improves consistency of individual sample labels but also leads to significant improvements in the classification performance of the task.

4.5 Evaluating Synthetic–Real Data

To evaluate the practical utility of synthetic samples, we measure classification performance when

Table 2: Classification accuracy of the fine-tuned SciBERT baseline, SciBERT prompting, and GPT-4o prompting on the Arxiv and WOS-11967 datasets. Results for open-source LLMs—LLaMA 3.2 1B (denoted as LLaMA-1B), LLaMA 3.2 3B (LLaMA-3B), and Phi-4-mini (Phi-4)—are also included for comparison.

SciBERT Baseline			
Dataset	**Accuracy**	**Dataset**	**Accuracy**
Arxiv	0.8828	WOS-11967	0.9005

Dataset	**Method**	**LLaMA-1B**	**LLaMA-3B**	**Phi-4**	**Zero Shot**	**k=2**	**k=3**	**k=5**
Arxiv	Prompt Based	–	–	–	0.78	0.85	0.86	0.88
	Base Synthetic	0.8864	0.8868	0.8824	–	–	–	–
	DPO Synthetic	0.8872	0.8876	0.8904	–	–	–	–
	U-DPO	0.8884	0.8896	**0.8912**	–	–	–	–
WOS	Prompt Based	–	–	–	0.64	0.78	0.82	0.85
	Base Synthetic	0.9076	0.9085	0.906	–	–	–	–
	DPO Synthetic	0.9118	0.9112	0.9122	–	–	–	–
	U-DPO	0.9143	**0.9156**	0.9143	–	–	–	–

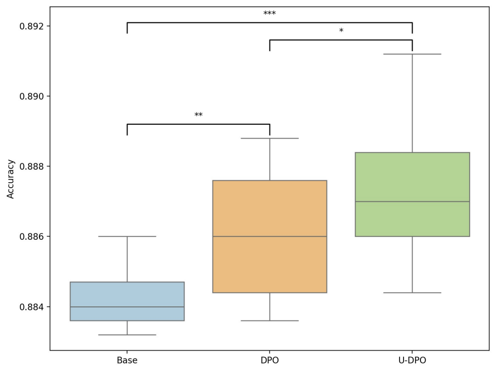

Figure 5: Classification accuracy on the Arxiv dataset when combining a fixed set of real data with synthetic samples generated using different methods. U-DPO achieves the highest performance among all hybrid setups.

combining them with a fixed subset of real annotated data. This setup reflects a realistic scenario in which synthetic augmentation is used to improve generalization. To examine the impact of synthetic data volume, we vary the number of generated samples per class across $\{10, 30, 50, 100, 150\}$. Interestingly, models trained with 50 synthetic samples per class consistently achieved the highest performance on average, suggesting that moderate augmentation achieves an effective balance between synthetic diversity and label reliability.

We conduct experiments using a hybrid training set composed of synthetic samples generated by three different methods: baseline generation, standard DPO, and U-DPO. Each synthetic set is combined with a fixed number of real samples per class.

As shown in Table 2, augmenting real data with synthetic samples leads to consistent accuracy improvements across both datasets. Among the methods, U-DPO yields the most substantial gains, indicating that utility-based optimization improves both the quality of the standalone sample and the downstream effectiveness in hybrid settings.

In addition, we compare our approach against GPT-4o (OpenAI, 2023) prompting baselines under zero-shot and few-shot conditions. Despite GPT-4o's strength as an instruction-following model, classifiers trained on U-DPO synthetic data outperform both prompting setups.

Furthermore, paired t-tests conducted on accuracies of 20 independent runs indicate that U-DPO consistently outperforms both the baseline and standard DPO in a statistically significant manner. As shown in Figure 5, U-DPO also produces more stable and higher accuracy distributions between trials, strengthening the robustness of utility-based training. Specifically, DPO and U-DPO significantly outperform the baseline model ($p < 0.01$ and $p < 0.001$, respectively), while U-DPO further shows a significant improvement over standard DPO ($p < 0.05$). These results highlight the value of incorporating utility signals not only in optimizing preference alignment, but also in generating practically useful training data for supervised learning.

4.6 LLM-based Evaluation with GPT-4.5

To assess the quality of the generated synthetic samples, we employ GPT-4.5 as an automated evaluator. Each sample is rated on a 0–5 scale based on

Statistic	Standard DPO	U-DPO
Mean	4.05	4.14
Median	4.50	4.50
Std Dev	1.09	0.89

Table 3: GPT-4.5-based evaluation of synthetic samples from the Standard DPO and U-DPO.

its relevance, fluency, and class alignment. A total of 132 synthetic samples were evaluated, comprising three samples per class from both the Arxiv and WOS-11967 datasets.

Table 3 reports the mean, median, and standard deviation of scores assigned to generations from the Standard DPO and the U-DPO incorporating U-DPO. The U-DPO achieves a higher average score (4.14 vs. 4.05) and the same median score (4.50), indicating improved overall quality without compromising peak performance. Additionally, the lower standard deviation (0.89 vs. 1.09) suggests that the U-DPO produces more consistent outputs across samples.

These results suggest that utility-based generation improves not only the average quality but also the reliability of synthetic samples as judged by a strong LLM evaluator.

4.7 Expanded Discussion of Experimental Findings

The experimental results clearly demonstrate that U-DPO yields superior text classification performance compared to both standard DPO and the baseline synthetic data approach. Classifiers trained on U-DPO-generated synthetic datasets consistently outperformed those trained on either baseline synthetic text or text from a standard DPO-tuned model. This trend holds across both benchmark datasets, including WOS-11967 and Arxiv, as well as various model configurations, underscoring the robustness of U-DPO's improvements. Overall, U-DPO enhances classification accuracy by producing higher-quality synthetic data that better aligns with true labels, leading to more effective downstream performance across diverse datasets. Representative examples of the synthetic data generated under different training regimes are provided in the Appendix C for further reference. Considering the degree of training improvement, the LLaMA-3.2-1B model generally exhibited lower performance gains compared to larger models, suggesting that model size may play a significant role in the effectiveness of utility-based preference optimization. This observation is further supported by margin score evaluations and classification performance assessments conducted using only synthetic data, both of which indicate that larger models tend to produce higher-quality, better-aligned samples that translate into improved downstream results.

5 Conclusions

In this paper, we explored how preference-guided generation with large language models can improve the quality of synthetic data for text classification. We introduced Utility DPO (U-DPO), a variant of Direct Preference Optimization designed specifically for class-conditional generation. By incorporating a utility signal that promotes label-consistent and discriminative outputs, U-DPO produces synthetic examples that better reflect the needs of a classifier.

Our experiments on multiclass document classification show clear benefits: models trained on U-DPO-generated data consistently outperform those using baseline LLM outputs or standard preference tuning. Notably, we observed stronger accuracy and generalization to real test data. Even in low-resource scenarios, augmenting limited real examples with U-DPO samples led to substantial improvements.

A closer analysis using a margin-based metric revealed that U-DPO enhances label fidelity in generated texts, shedding light on why its samples are more effective for training.

Taken together, these results highlight the value of task-specific preference optimization in generating high-quality synthetic data. We believe this approach offers a practical and scalable way to reduce reliance on large annotated datasets, and we hope it encourages further exploration of preference-driven generation in NLP.

6 Limitation

Although our findings demonstrate the potential of U-DPO, several limitations remain. First, the effectiveness of our method depends heavily on the quality of the preference signal. In our case, preferences are derived from an automatic classifier using margin scores and class consistency checks, which may introduce biases or errors. This can lead to overoptimization of proxy metrics without real improvements in downstream performance.

Second, our evaluation is limited to two datasets in the domain of research article classification. It remains to be seen whether U-DPO generalizes to other tasks such as short-text or multilabel classification, or to domains where discrete class labels are not clearly defined.

Finally, preference-based fine-tuning introduces computational overhead. Although we used relatively lightweight models (up to 4B parameters), scaling to larger models or fine-grained label spaces may be prohibitively expensive due to the cost of pairwise comparisons.

In sum, U-DPO offers a promising direction for improving synthetic data quality, but further work is needed to refine the preference signal, reduce noise, and evaluate generalization across tasks and domains.

7 Future work

There are several promising directions for extending our utility-based synthetic data generation framework.

First, improving the quality of the preference signal is key. Instead of relying solely on margin-based scores, future work could explore automatic preference inference through model-internal scoring, ensemble agreement, or task-specific heuristics, that better reflect sample utility.

Second, testing U-DPO on more diverse data types is crucial to validate its generality. Applications to short-form texts, multilabel tasks, noisy labels, or domain-specific corpora would show how well the method adapts to varied real-world settings.

Third, reducing the overhead of margin-based selection is an important efficiency challenge. Since utility scoring requires repeated model evaluations, scalable alternatives such as ranking distillation, selective pair mining or joint training with the classifier may improve both speed and quality.

Finally, analyzing the generated data itself can offer insight into what U-DPO learns. Understanding linguistic patterns, diversity, and preference-driven behaviors could guide future improvements in synthetic supervision strategies.

Looking ahead, these directions point to a more efficient, adaptable, and purposeful use of LLMs for task-specific data creation.

8 References

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *Preprint*, arXiv:2406.08660.

Duo Chai, Wei Wu, Qinghong Han, Fei Wu, and Jiwei Li. 2020. Description based text classification with reinforcement learning. *Preprint*, arXiv:2002.03067.

Juhwan Choi, Yeonghwa Kim, Seunguk Yu, JungMin Yun, and YoungBin Kim. 2024. UniGen: Universal domain generalization for sentiment classification via zero-shot dataset generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1–14, Miami, Florida, USA. Association for Computational Linguistics.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset. *Preprint*, arXiv:1905.00075.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? *Preprint*, arXiv:2403.17661.

Zeyu Gan and Yong Liu. 2025. Towards a theoretical understanding of synthetic data in llm post-training: A reverse-bottleneck perspective. *Preprint*, arXiv:2410.01720.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *Preprint*, arXiv:2501.08457.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th*

IEEE International Conference on Machine Learning and Applications (ICMLA), page 364–371. IEEE.

Udo Kruschwitz and Maximilian Schmidhuber. 2024. LLM-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 37–51, Torino, Italia. ELRA and ICCL.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *Preprint*, arXiv:2310.07849.

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. *Preprint*, arXiv:1907.11692.

Hamed Meshkin, Joel Zirkle, Ghazal Arabidarrehdor, Anik Chaturbedi, Shilpa Chakravartula, John Mann, Bradlee Thrasher, and Zhihua Li. 2024. Harnessing large language models' zero-shot and few-shot learning capabilities for regulatory research. *Briefings in Bioinformatics*, 25(5):bbae354.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *Preprint*, arXiv:2503.14023.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Letian Peng, Zilong Wang, and Jingbo Shang. 2024. Incubating text classifiers following user instruction with nothing but LLM. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3753–3766, Miami, Florida, USA. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Are large language models good classifiers? a study on edit intent classification in scientific document revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15049–15067, Miami, Florida, USA. Association for Computational Linguistics.

Rohit Sharma, Shanu Kumar, and Avinash Kumar. 2025. Read: Reinforcement-based adversarial learning for text classification with limited labeled data. *Preprint*, arXiv:2501.08035.

Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. 2025. Direct multi-turn preference optimization for language agents. *Preprint*, arXiv:2406.14868.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Sowmya Vajjala and Shwetali Shimangaud. 2025. Text classification in the llm era – where do we stand? *Preprint*, arXiv:2502.11830.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *Preprint*, arXiv:2312.01044.

Zhiqiang Wang, Yiran Pang, Yanbin Lin, and Xingquan Zhu. 2024. Adaptable and reliable text classification using large language models. *Preprint*, arXiv:2405.10523.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. 2024. Cal-dpo: Calibrated direct preference optimization for language model alignment. *Preprint*, arXiv:2412.14516.

Yosuke Yamagishi and Yuta Nakamura. 2024. UTRad-NLP at #SMM4H 2024: Why LLM-generated texts fail to improve text classification models. In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 42–47, Bangkok, Thailand. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linda Zeng. 2024. Leveraging large language models for code-mixed data augmentation in sentiment analysis. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 85–101, Miami, Florida, USA. Association for Computational Linguistics.

Prompt for Synthetic Data

You are an expert academic assistant. The following Examples are academic paper abstracts in the field of text classification and synthetic data generation. They are written in formal scientific style.

You are an expert academic assistant. The following Examples are academic paper abstracts in the field of something. They are written in formal scientific style. Your task is to generate a new academic abstract in a similar style and topic.

Examples:

- **Example 1:**
  {abstract 1}

- **Example 2:**
  {abstract 2}

Now, generate a single academic abstract paragraph in the same domain.
Only output the abstract content. Do not include titles, citations, links, or additional instructions. Abstract:

Abstract:

A Additional Margin Score Results

Figures 6 and 7 illustrate the distribution of margin scores for models such as `LLaMA 3.2 1B` and `LLaMA 3.2 3B` on datasets in Arxiv. Consistently across these combinations, both standard DPO and our proposed Utility DPO (U-DPO) training achieve higher median and average margin scores relative to baseline generation without preference optimization.

While some labels exhibit slightly lower margin scores compared to the `Phi-4-mini` model, the majority demonstrate improvements, confirming

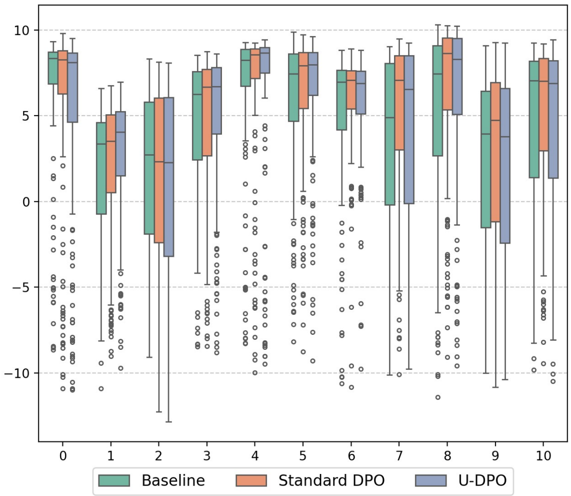

Figure 6: Margin score distributions for synthetic samples generated from identical prompts using the `LlaMA3.2-1B` model on the `Arxiv` dataset.

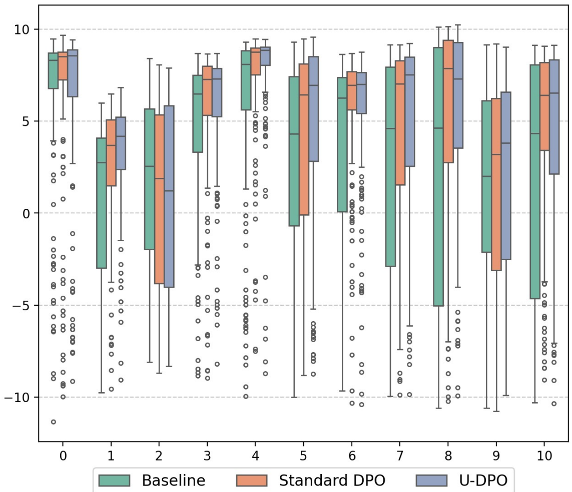

Figure 7: Margin score distributions for synthetic samples generated from identical prompts using the `LlaMA3.2-3B` model on the `Arxiv` dataset.

the general effectiveness of preference-based optimization across diverse model and dataset configurations.This suggests that model size and architecture play significant roles in the effectiveness of preference-guided training, with smaller models like `LLaMA 3.2 1B` benefiting less from the DPO and U-DPO approaches relative to larger or more optimized models.

B Additional Results on Classification Performance of Synthetic Data

Figures 8 report classification results for models trained exclusively on synthetic samples generated from the WOS-11967 dataset. We compare three generation methods: base model sampling, stan-

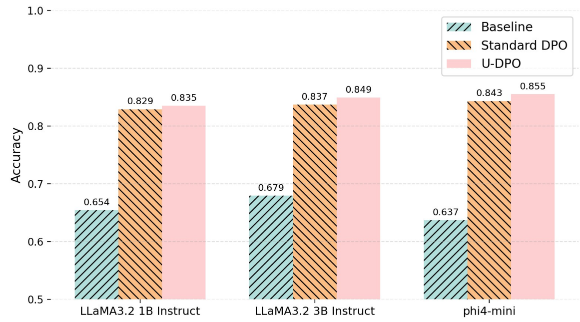

Figure 8: Downstream classification accuracy of models trained exclusively on synthetic data generated under three training regimes: All models are evaluated on the WOS-11967 test set.

dard DPO, and Utility DPO (U-DPO). Consistent with observations on Arxiv data, classifiers trained on U-DPO synthetic samples exhibit superior accuracy compared to those trained on samples from the base model or standard DPO.

C Generated Sample Examples

Table 4 shows representative synthetic samples generated by the Standard DPO, and U-DPO for a specific class label. These examples illustrate differences in content quality, fluency, and label alignment.

Method	**Generated Sample (Class: Computer Vision)**
Standard DPO	In recent years, the integration of computer vision and computer graphics technologies has led to significant advancements in the development of interactive virtual environments. This paper discusses the challenges and solutions related to the seamless integration of virtual objects into real-world images. We explore the importance of accurate pose estimation, surface correspondence, and the creation of visually convincing interactions between virtual and real-world elements. Our study introduces novel algorithms for the accurate placement and interaction of virtual objects within complex environments. We demonstrate the efficacy of these algorithms through a series of case studies, including architectural visualization and interactive gaming scenarios. Our findings suggest that advancements in this area have far-reaching implications for the fields of computer graphics, computer vision, and augmented reality applications, providing users with more immersive and realistic experiences.
U-DPO	This study presents a novel framework for integrating visual-inertial odometry with structure-from-motion algorithms to achieve real-time 6-DOF pose estimation and mapping. Our proposed method leverages the complementary strengths of visual and inertial sensors to mitigate the limitations inherent in each modality. By employing a robust factor graph optimization, we ensure accurate alignment of visual and inertial data, which is crucial for achieving precise motion tracking and environmental mapping. Experiments conducted in diverse environments, including indoor and outdoor settings, demonstrate our approach's resilience to challenging conditions such as variable lighting and dynamic occlusions. The results indicate significant improvements in accuracy and robustness compared to traditional visual-inertial odometry systems, paving the way for enhanced autonomous navigation and interactive robotic applications. Future research directions include refining the sensor fusion algorithm to further reduce computational overhead and exploring the integration of additional sensory modalities, such as LiDAR, to complement the visual-inertial system.

Table 4: Example synthetic samples generated by different methods for the same class.

Method	**Generated Sample (Class: Materials Science, Engineering)**
Standard DPO	The study of process optimization in manufacturing engineering involves determining the effect of various process parameters on product quality, cost, and efficiency. Recent advancements in statistical and computational methods have provided engineers with powerful tools to design experiments that can systematically investigate these effects. In this paper, we present a comprehensive analysis of the application of Design of Experiments (DoE) and Taguchi Methods in optimizing manufacturing processes. We demonstrate how these methodologies can be effectively utilized to identify critical factors, interactions, and optimal settings for process variables. Our case study focuses on the manufacturing of a high-precision component, where we employed a full factorial design and Taguchi orthogonal arrays to minimize variation and improve product quality. The results highlight the potential of these approaches to enhance process control, reduce waste, and achieve cost-effective production. Further, we discuss the integration of these techniques with modern manufacturing systems and the implications for continuous improvement and innovation in engineering practice.
U-DPO	In this study, we examine the integration of piezoelectric sensors into composite manufacturing processes. Piezoelectric sensors have emerged as a valuable tool for monitoring stress distribution and detecting potential failure points within composite materials. This paper presents a comprehensive analysis of the impact of piezoelectric sensor embedding on the mechanical properties of composite materials, particularly focusing on tensile strength and strain distribution. Our results indicate a slight reduction in tensile strength but an improved ability to monitor stress distribution, which can lead to better quality control and increased safety in composite manufacturing. Additionally, the study explores the feasibility of using piezoelectric sensors as a real-time monitoring system, offering insights into their potential applications in aerospace and automotive industries. The findings underscore the importance of piezoelectric sensors in enhancing the reliability and performance of composite materials used in critical applications.

Table 5: Example synthetic samples generated by different methods for the same class.