Use Random Selection for Now: Investigation of Few-Shot Selection Strategies in LLM-based Text Augmentation

Anonymous ACL submission

Abstract

The generative large language models (LLMs) are increasingly used for data augmentation tasks, where text samples are paraphrased (or generated anew) and then used for downstream model fine-tuning. This is useful, especially for low-resource settings. For better augmentations, LLMs are prompted with examples (few-shot scenarios). Yet, the samples are mostly selected randomly, and a comprehensive overview of the effects of other (more "informed") sample selection strategies is lacking. In this work, we compare sample selection strategies existing in the few-shot learning literature and investigate their effects in LLM-based textual augmentation in a low-resource setting. We evaluate this on in-distribution and out-of-distribution model performance. Results indicate that while some "informed" selection strategies increase the performance of models, especially for out-of-distribution data, it happens only seldom and with marginal performance increases. Unless further advances are made, a default of random sample selection remains a good option for augmentation practitioners.

1 Introduction

The emergence of recent large language models (LLMs) such as GPT-4, Gemini, Llama, and their wide availability prompted their use in *augmentation* of textual datasets (Ubani et al., 2023; Dai et al., 2023; Piedboeuf and Langlais, 2023; Cegin et al., 2023, 2024a). LLM augmentation has been used in various domains such as sentiment analysis (Onan, 2023; Piedboeuf and Langlais, 2023), intent classification (Cegin et al., 2023), news classification (Piedboeuf and Langlais, 2023; Cegin et al., 2024a), and health symptoms detection (Dai et al., 2023). These augmentations are often performed in a low-resource setting with a limited number of seed samples. In most LLM-based augmentation scenarios, the dataset size is increased through *paraphrasing* of original samples or *generation* of completely new samples that adhere to a specified label. This can be done without any samples provided (*zero-shot*). Alternatively, one can include already existing samples as part of the prompt to better instruct the LLM (*few-shot*). The augmented datasets are then used for training *downstream* models, which are usually much smaller than the prompted LLMs, and thus cheaper and more suitable for production environments.

Recent studies report better performance for few-shot LLM-based augmentation, as compared with zero-shot approaches (Cegin et al., 2024a; Piedboeuf and Langlais, 2024). Most existing few-shot augmentation studies select the samples randomly, and the potential of using more *informed* selection strategies (existing elsewhere in few-shot learning literature) is under-explored. Furthermore, augmentation studies focus only on paraphrasing and are evaluated on in-distribution data.

In few-shot learning, the *informed* sample selection strategies aim to select the most relevant samples that would lead to better outputs. The samples can be selected based on their similarity, diversity, informativeness, or quality (Li and Qiu, 2023; Zhang et al., 2022; Chang and Jia, 2023; Pecher et al., 2024b). Through these methods, LLMs can potentially produce better augmentations in return for the additional computation costs of the informed sample selection. *Literature shows that the choice of samples for few-shot learning significantly influences its outcomes* (i.e., sensitivity of sample selection) (Pecher et al., 2024a; Zhang et al., 2022; Köksal et al., 2023; Agrawal et al., 2023). For example, recent studies have investigated the effects of such sample selection strategies for in-context learning (Zhang et al., 2022; Li and Qiu, 2023) or LLM alignment (Zhou et al., 2024). However, *for augmentation scenarios, an investigation of sample selection strategies effects is lacking*.

The goal of this paper is to compare existing

sample selection strategies in few-shot text augmentation for a low-resource setting. This comparison is measured by the performance of downstream models trained on the augmented data. We investigate the typical *paraphrasing* scenario, but also less covered *generation of new samples*. Along with more frequent in-distribution (ID), we also evaluate out-of-distribution (OOD) data. We run our experiments for various LLMs and tasks. We identify the best-performing sample selection strategy in each scenario (parameter combination) and compare it against two baselines: (1) the zero-shot augmentation and (2) the few-shot augmentation with random sample selection. We formulate the following research questions:

RQ1: *Considering downstream model performance, which sample selection strategy performs the best most consistently? (when considering both in-distribution and out-of-distribution setups).*

RQ2: *Considering downstream model performance, when and how often do the best-performing sample selection strategies outperform the baseline strategies?*

We compared 8 different sample selection strategies (see 3.1) against 2 baseline strategies (*zero-shot/no-samples* strategy and *random samples* strategy) on 3 different LLMs (Llama-3.1, Mistral-v0.3, and Gemma-2). We experimented with 8 different datasets (for sentiment analysis, news classification, question topic classification, paraphrase detection, and natural language inference) with both in-distribution and out-of-distribution splits on RoBERTa as our downstream model. We used a low-resource setting, using only 20 samples per label. Furthermore, we also investigated the *composition of the examples* from the point of labels (whether it is more beneficial to include samples only from the target label being augmented or also from other labels). We investigated two augmentation techniques: *paraphrasing* of samples and *generation* of completely new samples. We repeated the whole process 3 times with different random seeds, ensuring the robustness of our results.

The most prominent findings are: 1) None of the existing sample selection strategies is consistently better than the baseline in the majority of cases for in-distribution, 2) Selecting examples at random yields the best performance in the majority of cases and does not require additional overhead, 3) For out-of-distribution, the *synthetic samples dissimilarity* selection strategy yields the highest performance more often than the baseline strategies. It can be considered for uses where overhead selection costs are not an issue.

2 Related Work: LLM-based Text Augmentation

Soon after their advent, new LLMs, such as GPT-4 or Llama, started to be used as data augmentation tools, leveraging their ability to produce a diversity of texts. The LLM-based augmentation is typically done through paraphrasing (Cegin et al., 2024a; Dai et al., 2023; Sen et al., 2023). Less often, LLMs are used to create semantically new samples adhering to a given label (Ubani et al., 2023). LLM-based augmentation has been used for a variety of augmentation tasks such as automated scoring (Fang et al., 2023), low-resource language generation (Ghosh et al., 2023), intent classification (Sahu et al., 2022), sentiment analysis (Piedboeuf and Langlais, 2023; Ubani et al., 2023; Onan, 2023; Yoo et al., 2021), hate speech detection (Sen et al., 2023), news classification (Piedboeuf and Langlais, 2023), content recommendation (Liu et al., 2024), and health symptoms classifications (Dai et al., 2023).

Recent studies have also used few-shot learning as part of the augmentation by supplying the LLM with various examples from the dataset in the prompts. It has been leveraged for named entity recognition (Ye et al., 2024), classification performance (Cegin et al., 2024a) or text summarization (Sahu and Laradji, 2024). While the performance of the few-shot approaches in augmentation seems to outperform zero-shot ones (where no examples are used) (Piedboeuf and Langlais, 2024), the effects of various sample selection strategies are under-explored, as many studies simply select the samples randomly. Only one study explored other strategies (Cegin et al., 2024a), which used a human-inspired sample selection strategy.

While sample selection strategies have found their usage in various in-context learning tasks (significantly altering the performance of LLMs) and while some studies already hint at increased performance of few-shot augmentation over zero-shot augmentation (Cegin et al., 2024a; Piedboeuf and Langlais, 2024), an investigation of various sample selection strategies for LLM-based augmentation methods is completely lacking.

3 Study Design

To assess which sample selection strategies work best for LLM-based data augmentation, we performed a comparative study in a low-resource setting. The same basic scenario was used in each case: given a dataset, 20 seed samples were selected from each label. For each seed sample, a given LLM "augmented" the samples 5 times. This was repeated for each sample selection strategy and type of augmentation technique used (paraphrasing or creating completely new samples). Next, a downstream model was fine-tuned on both sub-sampled data and augmented samples and then evaluated on in-distribution and out-of-distribution data. For in-distribution data, we used the original test splits of each dataset, while for the out-of-distribution data, we used test splits from a different dataset with the same task (e.g. *Yelp* dataset test split was used as out-of-distribution data when evaluating performance on the *Tweet Eval* dataset for sentiment analysis).

This scenario was repeated for all sample selection strategies and baselines for a variety of parameters (see below). Then, the performance of the models (measured by F1-macro) was compared for each sample selection strategy to answer the RQ1. This was followed by comparing the best-performing sample selection strategies against the best-performing baseline strategy of either zero-shot (no examples provided) or randomly selected examples to answer RQ2. We publish all of our results, the code, and the data used [1].

We used a broad range of study parameters to ensure the robustness of our results by using both the baseline strategies and the sample selection strategies in a variety of cases. We include the different augmentation techniques and example compositions in terms of labels to capture a wide variety of cases. The whole process was repeated 3 times, and different seed samples were selected. The study had the following parameters:

- 8 sample selection strategies (Forgetting with 2 variations, Cartography with 3 variations, Cosine similarity/dissimilarity and Synthetic samples dissimilarity) with 2 baseline strategies (zero-shot with no examples provided and random few-shot with examples selected randomly),
- 3 LLMs used as augmenters (LLama-3.1-8B, Gemma-2-9B and Mistral-v0.3-7B),
- 8 datasets used (*MNLI, QQP, Yelp, Tweet sentiment evaluation, AG News, News Topic, Yahoo, Trec*),
- 2 types of *composition of examples* used (examples used only from the target label or examples selected from all labels in the dataset),
- 2 augmentation techniques (either paraphrasing of existing samples or generation of new label adhering samples),

This resulted in 1,300 combinations for which downstream models were trained and evaluated repeatedly.

3.1 Sample Selection Strategies

We used the best-performing sample selection strategies identified by previous studies on sample selection in in-context learning (Pecher et al., 2024b; Li and Qiu, 2023; Chang and Jia, 2023; Toneva et al., 2018; Zhang and Plank, 2021). We used 5-shots per label for each of the sample selection strategies.

First, we used the *Similarity* and *Dissimilarity* selections that are currently the most popular selection strategies for in-context learning (An et al., 2023; Liu et al., 2022; Chang and Jia, 2023). To select the samples, we calculated the cosine similarity between the feature representation of the samples and then selected either the most similar or the most dissimilar ones. In the case of paraphrasing, we calculated the similarity of the sample we were augmenting. In the case of generation, we first randomly select one sample and then calculate the similarity of this sample.

Second, we used the *Synthetic samples dissimilarity* sample selection (Cegin et al., 2024a). To select the samples, we first use the LLM to generate a set of synthetic samples and then use the *dissimilarity* selection to select the set of examples from this set. This is different from the *Dissimilarity* above, as this method uses synthetic data, while the original uses data from the dataset itself.

Third, we used the *Forgetting* strategy that selects the samples based on how often they are forgotten (Toneva et al., 2018). To select the samples, we first trained the model on the underlying task for a fraction of the overall epochs and observed the training dynamics. For each sample, we calculated how often the prediction of the model was incorrect after it had already been correct in the previous

[1] Data and code in the ZIP file

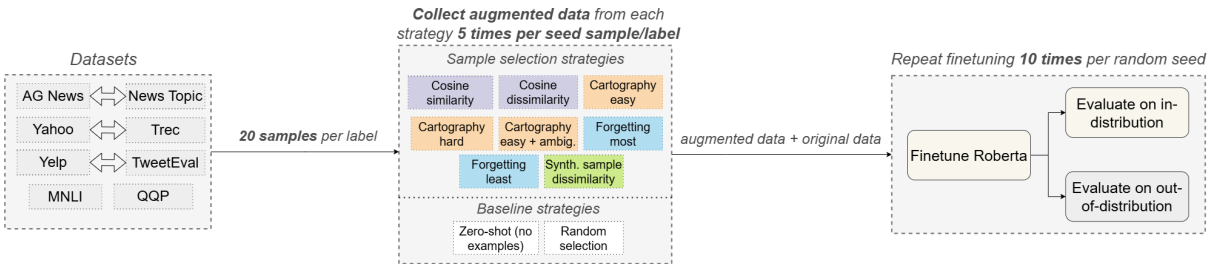

Figure 1: Overview of our methodology. For each dataset, we randomly sample 20 samples per label, which are then used to collect up to 5 augmented samples per seed sample. These seeds are used for fine-tuning with the augmented samples to evaluate each sample selection strategy. This entire process is repeated 3 times with different random seeds. Similar sample selection strategies have the same colour.

epoch. Afterward, these forgetting events are used to select the samples. We explored two different settings in our experiments and chose samples accordingly: 1) *Forgetting most*, where we selected the samples that were the most often forgotten; and 2) *Forgetting least*, where we selected the samples that were forgotten the least number of times.

Finally, we used the *Cartography* sample selection that measures how easy or hard it is to learn the different samples (Swayamdipta et al., 2020; Zhang and Plank, 2021). This ease of learning is determined by training the model on the underlying task for a fraction of the overall epochs and looking at the average confidence/probability of the correct predictions and the variance of this confidence. The samples with high confidence and low variance are considered to be the *easy* to learn samples. At the same time, the samples with small confidence and small to medium variance are considered the *hard* to learn ones. The remaining samples are considered to be *ambiguous* (medium confidence or samples with high variance). We explored three different settings in our experiments. We chose the samples accordingly: 1) **Easy** samples, where we sorted the samples based on confidence and choose the top 5 samples with highest confidence; 2) **Hard** samples, where we sorted the samples based on confidence and choose the bottom 5 samples (i.e., the lowest confidence samples); 3) **Easy + Ambiguous**, where we first calculated average confidence, selected the samples whose confidence is higher than the average, and then randomly sampled from them.

Additionally, we opted against using active learning methods: as observed by previous works on classification using ICL (Li and Qiu, 2023; Pecher et al., 2024c), as well as concurrent work on selecting samples for augmentation (Wang et al., 2025), the active learning strategies perform on par (but often worse) than the other selection strategies. For this reason, we have decided to forego the active learning methods in our experiments, as they often require more computation resources (due to their iterative selection of samples).

3.2 Datasets

For a diverse evaluation, we selected 8 datasets representing tasks of sentiment analysis, news classification, question topic classification, paraphrase detection, and natural language inference. We used the *News Category* (Misra, 2022; Misra and Grover, 2021) and *AG news* (Zhang et al., 2015a) for news classification, *Yahoo* (Zhang et al., 2015b) and *Trec* (Li and Roth, 2002) for question topic classification, MNLI dataset (Williams et al., 2018) for natural language inference, Quora Question Pairs Dataset (QQP) for paraphrase detection (Wang et al., 2017) and *TweetEval* (Rosenthal et al., 2017) and *Yelp* (Zhang et al., 2015a) for sentiment classification. All datasets were in English. For the in-distribution evaluation of models, we used the test split of each dataset. For out-of-distribution evaluation for each dataset, we used the test split of the dataset that is within the same domain, e.g., we used the test split from *Yelp* for *TweetEval* and vice versa (with the exception of MNLI, which has its own out-of-distribution test split and QQP, for which we used the PAWS (Zhang et al., 2019) dataset as out-of-distribution). While still of the same task, we considered these splits out-of-distribution due to them being collected from other domains or sources (e.g. sentiment analysis of Yelp reviews for models trained on tweets). We only generated/paraphrased hypotheses for MNLI, given the premise from the dataset. We also only generated/paraphrased one paraphrase for QQP and left the others intact. Details about labels used and preprocessing can be found in Appendix C.

3.3 Evaluation Process

We randomly selected 20 samples per label from each dataset and repeated this three times with different random seeds. We chose 20 samples per label as this number of seed samples per label should yield the highest effect for augmentation (Cegin et al., 2024b). We then augmented the entire selected subset of the dataset for each combination of augmentation technique (*paraphrasing* or *generation*), sample selection strategy (including baselines), augmenting LLM, and *composition of the examples* from the point of labels. We instructed the LLM to collect 5 new samples per seed sample for each combination of parameters. Prompt templates, specific versions of LLMs used, and parameters used for the LLMs can be found in Appendix D. We did not check the validity of the collected samples, as previous works have already shown that the validity of LLM augmentation methods is quite high (Cegin et al., 2023, 2024a).

We used RoBERTa-base for fine-tuning and used the version of the model from Huggingface. The best working hyperparameters were found via hyperparameter search, and these can be found in Appendix B. We trained each model 10 times per each random seed and augmentation parameter combination. The models were trained separately on the data collected from Llama-3.1, Gemma-2, and Mistral. Finally, we computed the F1-macro of all fine-tuned models to allow the comparison of sample selection strategies between themselves and against the baseline strategies.

4 Study Results

Our study has multiple parameter dimensions, which yielded more than 1,300 combinations. We aggregated the results for each of the used LLMs. During our analysis, we did not identify any LLM bias towards one of the sample selection strategies, as the 3 used LLMs performed similarly.

To keep the comparison of various sample selection strategies simple, we only compare the best-performing sample selection strategy combination on the dataset given the augmentation techniques of either *generation* or *paraphrasing* and *composition of labels* in terms of labels. We also use the same setting for the baseline strategies of zero-shot and random few-shot. We wish to identify strategies that provide the best performance most consistently (in most cases) and outperform the baselines the most. We analyze the different augmentation techniques and *composition of labels* and how they influence the model performance in Appendix F.

We distinguish between the best-performing sample selection strategy for in-distribution data and out-of-distribution data for each of the datasets. To identify the best-performing sample selection strategy (including the baselines) in these cases, we compute the mean of the model performance across all of the random seeds and compare these means. There were a total of 72 cases for 8 datasets, 3 different LLMs, and 3 different random seeds used. After identifying the best-performing sample selection strategy, we statistically tested its distribution of model performance against the best-performing baseline strategy (either zero-shot or random few-shot based on their mean) using Mann-Whitney-U tests with $p=0.05$ to measure the number of times the sample selection strategies are statistically significantly better than the best baseline strategy.

4.1 Best Performing Sample Selection Strategies

The number of times where each sample selection strategy (including baseline strategies) performed the best for each dataset for in-distribution (ID) and out-of-distribution (OOD) data can be found in Table 1. The comparison excluding baseline strategies can be found in Appendix E, together with the performance distributions for each sample selection strategy and dataset. **There is no apparent strategy that performed the best across all datasets for both in-distribution and OOD model performance.** However, certain sample selection strategies did perform best overall for given data distributions - the *Cartography with easy and ambiguous samples* performed the best most often from all sample selection strategies (excluding baseline strategies) for in-distribution data in 11 out of 72 cases (15.28%) and the *Synthetic samples dissimilarity* performed the best most often for OOD data in 23 out of 72 cases (31.94%).

Some of the strategies seem biased for certain datasets, performing well in those cases. For example, the *Synthetic samples dissimilarity* strategy is well suited for the MNLI dataset for both in-distribution and OOD cases.

Considering the sample selection strategies without the baselines, the *Cartography eas.+ambig. samples* and *Forgetting least* strategies perform best for in-distribution data, with both of them achieving the best performance in 13 out of 72 cases (18.06%). For OOD comparison of strate-

DATASET→ / Strategy↓	AGNEWS ID	AGNEWS OOD	NTOPIC ID	NTOPIC OOD	YAHOO ID	YAHOO OOD	TREC ID	TREC OOD	TEVAL ID	TEVAL OOD	YELP ID	YELP OOD	MNLI ID	MNLI OOD	QQP ID	QQP OOD	**TOTAL** ID	**TOTAL** OOD
Zero-shot	0	0	0	0	1	1	0	**2**	0	1	0	1	1	1	0	1	2	7
Random	1	0	**4**	1	1	0	**3**	0	**2**	1	**4**	0	1	1	0	0	**16**	8
Cos. sim.	0	0	1	**3**	1	1	1	0	0	0	0	0	0	0	1	0	4	4
Cos. dis.	0	1	0	0	0	1	0	1	1	0	0	**2**	1	1	0	0	6	6
Forget. most	**2**	0	1	2	0	**3**	0	1	1	1	0	0	0	0	1	0	5	7
Forget. least	**2**	3	1	0	0	0	**3**	0	**2**	1	0	1	0	0	1	1	9	6
Carto. hard	1	0	0	1	**3**	2	0	0	0	0	1	**2**	0	0	1	1	6	6
Carto. e.+amb.	0	1	0	0	2	1	0	**2**	**2**	1	3	1	1	1	**3**	1	11	8
Carto. easy	1	0	1	0	1	0	1	1	1	0	1	0	0	0	1	1	7	2
Synth. dis.	**2**	**4**	1	2	0	0	1	**2**	0	**4**	0	**2**	**5**	**5**	1	**4**	10	**23**

Table 1: No. cases for each sample selection strategy, including baseline strategies, where each strategy performed the best for each dataset for in-distribution (ID) and out-of-distribution (OOD) data. The last *Total* column aggregated all cases for that specific strategy. In total, only the *Synthetic samples dissimilarity* strategy on out-of-distribution outperforms the baseline strategies most often, while the random few-shot baseline strategy works best for in-distribution.

gies, the best strategy is the *Synthetic samples dis.* method in 28 out of 72 cases (38.89%) followed by the *Cartography eas.+ambig. samples* strategy in 9 out of 72 cases (12.5%).

We answer the *RQ1* as follows: Considering the sample selection strategies without the baseline strategies, the most effective sample selection strategy is *Cartography eas.+ambig. samples* for in-distribution and *Synthetic samples dis.* method for OOD. However, we also note that in certain cases, both of these strategies fail to perform as the best strategy even once (e.g. *NewsTopic* for *Cartography eas.+ambig. samples* and *Yahoo* for *Synthetic samples dis.* method).

4.2 Comparison of Best Sample Selection Strategies Against Baseline Strategies

We compare the best-identified sample selection strategies from Section 4.1 against baseline strategies as per Table 1 and also provide aggregated difference across all cases in mean F1-Macro for various sample selection strategies against the best-performing baseline of either random few-shot or zero-shot in Figure 2.

For in-distribution classifier performance, we identified as the best-performing sample selection strategy the *Cartography with easy and ambiguous samples* performing best in 11 out of 72 cases (15.28%). The best-performing baseline on in-distribution data is random few-shot, which achieved the best performance in 16 out of 72 cases (22.22%), an increase of 5 cases compared to the *Cartography eas.+ambig. samples* strategy. Out of the 11 cases where the *Cartography eas.+ambig. samples* performed best, it was statistically significantly better than the best baseline strategy in 7 cases (63.63%). The random few-shot baseline also achieved the best performance in a variety of cases across all the datasets, which the *Cartography eas.+ambig. samples* strategy did not and was outperformed by the *Cartography eas.+ambig. samples* strategy only on the *Yahoo* and *QQP* datasets.

For OOD classifier performance, we identified as the best-performing sample selection strategy the *Synthetic samples dissimilarity* performing best in 23 out of 72 cases (22.22%). The best-performing baseline on OOD data is random few-shot, which performed best in 8 out of 72 cases (11.11%), performing worse than the *Synthetic samples dis.* in 15 cases. Out of the 23 cases where the *Synthetic samples dis.* performed best, it was statistically significantly better than the best baseline strategy in only 6 out of the 23 cases (26.09%). In comparison, the *Synthetic samples dis.* works well on most datasets (as it achieves no best cases for *NewsTopic* and *Yahoo* datasets); the same can not be said about both the baselines: zero-shot strategy achieves no best cases on two datasets and random few-shot achieves no best cases on five datasets. However, these positive occurrences are hindered by only a few cases where the impact on performance is also statistically significant.

All of the sample selection strategies fail to make a consistent impact on model performance over the baselines, as can be seen in Figure 2. While there are cases where increases are apparent in both in-distribution and OOD performance (on the *MNLI* dataset), the sample selection strategies fail to outperform consistently the best baseline of either zero-shot with no examples or randomly

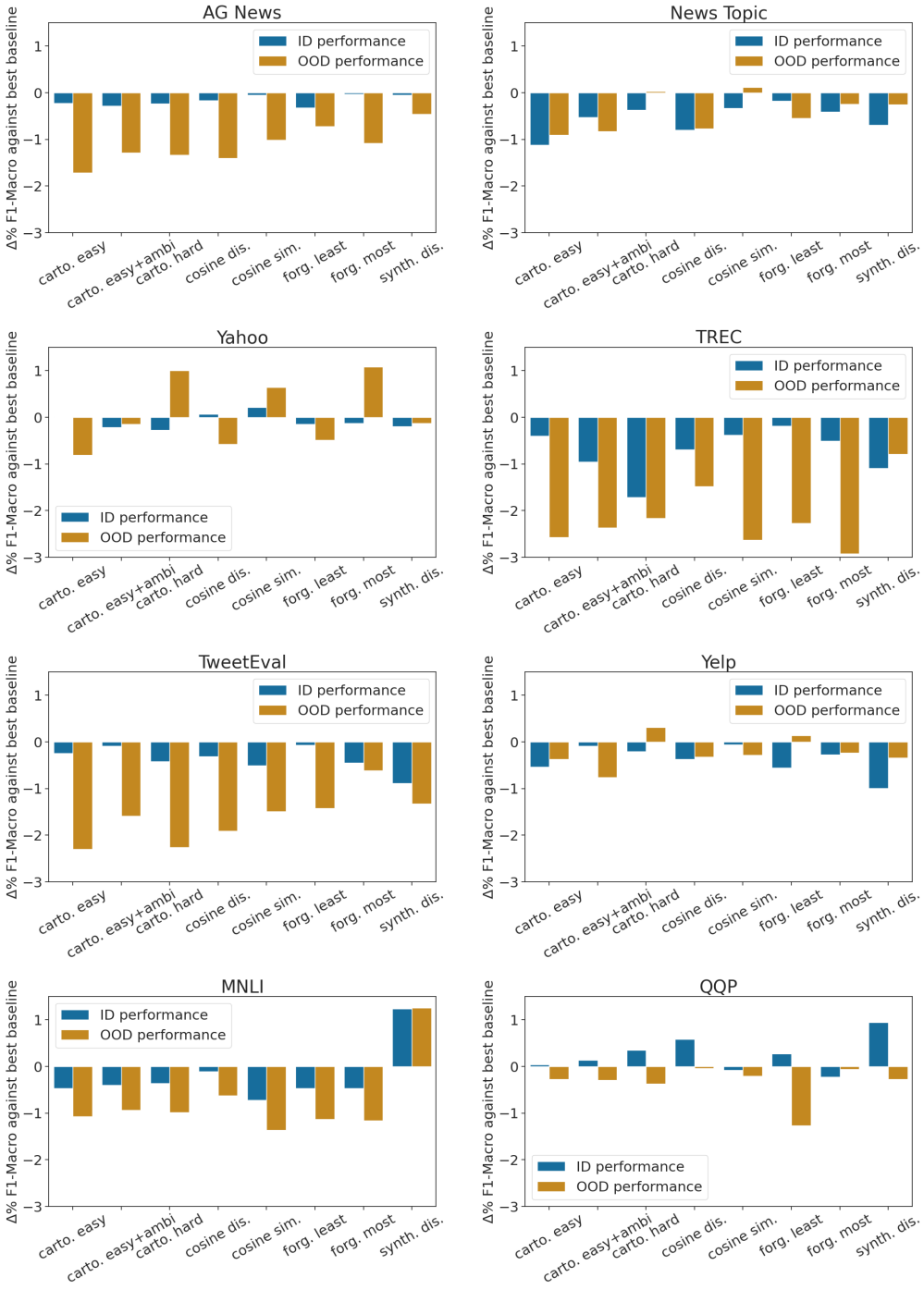

Figure 2: Aggregated difference across all LLMs and random seeds in mean F1-Macro for models trained on various sample selection strategies against the best-performing baseline of either random few-shot or zero-shot. While some strategies perform well in certain cases, as per Table 1, they fail to make a positive impact on model performance against baseline strategies in general.

selected samples for few-shot, as the increase in performance on one random seed is mitigated by losses of performance on another random seed with different seed samples.

We answer the *RQ2* as follows: When comparing in-distribution performance, the best baseline of random few-shot strategy performs better than the best sample selection strategy of *Cartography eas.+ambig. samples* in 5 more cases. Additionally, the random few-shot strategy works well across nearly all datasets for in-distribution classifier performance. When comparing OOD classifier performance, the best sample selection strategy of *Synthetic samples dis.* method performs better than

the best baseline of random few-shot in 15 more cases. However, most of these increases are not statistically significant, and the aggregated model performance in Figure 2 shows little to no benefit in most cases. Neither the baselines nor the *Synthetic samples dis.* method performs well on all datasets.

5 Discussion

The results of our experiments lead to the following observations: First, the *Cartography eas.+ambig. samples* strategy was best among sample selection strategies for in-distribution classifier performance. Such selection strategies seem to influence LLMs for text generation in similar ways as they do models fine-tuned on such selected data - they increase their performance on in-distribution data (Zhang and Plank, 2021).

Second, the *Synthetic samples dis.* strategy was best among sample selection strategies for OOD model performance. The strategy used in (Cegin et al., 2024a) was inspired by crowdsourcing (Larson et al., 2020) methods for collecting data for better OOD performance. This method appears to force LLMs to create more diverse samples by leveraging outlier synthetic data as examples, making the downstream models more robust.

Third, when comparing the best sample selection strategies against baseline strategies, the *Cartography eas.+ambig. samples* strategy does not outperform the random few-shot selection strategy. Not only does the random few-shot strategy perform better more often, but it does so more consistently across multiple datasets. In contrast, the *Cartography eas.+ambig. samples* strategy fails to perform best even once for some datasets. This hinders the applicability of this method, where it is clearly outperformed in some cases by baseline strategies or other sample selection strategies.

Fourth, the *Synthetic samples dis.* strategy outperforms the baseline strategies for OOD performance, but not across all datasets. However, neither the random few-selection selection strategy nor the zero-shot approach performs well on all datasets. This implies that increasing performance across all OOD cases for all datasets is a difficult problem. Additionally, the *Synthetic samples dis.* method is expensive, as it requires one additional inference from the LLM to select examples from.

Fifth, as seen in Figure 2, the aggregated increase of classifier performance when using sample selection strategies is small or negative, indicating that sample selection strategies do not work well for all random seeds. Given the increased costs of using sample selection strategies, this result favours the baseline strategies for text augmentation in general.

Sixth, comparing the baseline strategies between themselves, the random few-shot selection performs the best on in-distribution classifier performance. In contrast, the zero-shot strategy only performs well on OOD classifier performance. This might be due to the LLMs getting biased towards the examples provided and thus being more likely to produce augmentations that follow the distribution of the seed samples more closely. However, this might not be robust enough for good OOD classifier performance.

To summarise, while the *Synthetic samples dis.* strategy outperforms the baseline strategies for out-of-distribution classifier performance, the baseline strategies outperform the sample selection strategies for in-distribution classifier performance. However, any increase in classifier performance for both in-distribution and out-of-distribution is marginal and increases costs for collecting text augmentations. **While sample selection strategies work best in some cases, they do not so consistently. This underlines the need for better sample selection strategies for LLM-based text augmentation**.

6 Conclusion

We compared the effects of prominent sample selection strategies of few-shot learning for LLM-based text augmentation scenarios in a low-resource setting. We evaluated the downstream model performance on in-distribution and out-of-distribution data. We compared selection strategies against 2 baseline strategies (random few-shot and zero-shot). This comparison was done using 3 different LLMs, 8 different datasets, and 2 augmentation techniques (paraphrases and new samples).

Our comparison indicates that the baseline strategies outperform sample selection strategies for in-distribution performance. For out-of-distribution performance, the *Synthetic sample dissimilarity* strategy is best in more cases than the baseline strategies. However, the improvements are marginal and are not present in all datasets. Given the increased computations needed to use these sample selection strategies and their lacklustre performance, the baseline strategies represent a good default for few-shot augmentation practitioners.

Limitations

We note several limitations to our work.

First, we only used datasets, augmentation methods, and LLMs for the English language and did not investigate cases of multi-lingual text augmentation.

Second, we did not use various patterns of prompts and followed those used in previous studies (Cegin et al., 2023; Larson et al., 2020). Different prompts could have effects on the quality of text augmentations, but they would also radically increase the size of this study, and thus, we decided to leave this for future work and focused on the simplest prompts possible.

Third, we did not use newer LLMs for downstream model fine-tuning via PEFT methods (e.g., fine-tuning of Llama-3 or Mistral using QLoRA). While such inclusion would strengthen our findings, we decided not to use these models as evaluation of these models is very costly and takes a long time due to their size, which results in them being mostly used with a small subset of the testing data (Chang and Jia, 2023; Li and Qiu, 2023; Gao et al., 2021; Köksal et al., 2023). This, in return, can lead to unintentionally cherry-picked results. We see the usage of such fine-tunings as the extension of our work left for future work.

Fourth, for the LLM augmentation methods, we used only Llama-3.1-8B, Mistral-v0.3-7B, and Gemma-2-9B. We did not use larger models (e.g., 70B versions) as their increased performance in text augmentation for model accuracy has been shown (Cegin et al., 2024a) to be not that significant when compared to variants of LLMs with fewer parameters, while the inference costs compared to these smaller models are much higher.

Fifth, we used 5-shots on 20 seeds per label selected on each dataset. While a bigger number of seeds and shots could have been used, we opted for smaller numbers to keep the study manageable and the cost of the study low. In addition, a previous study (Pecher et al., 2024b) found that sample selection is more impactful when choosing only a small set of samples, and using more samples does not necessarily lead to better results due to the limited context size of the models. Furthermore, obtaining larger annotated datasets (e.g., hundreds of samples per class) is not feasible for many domains in practice. As such, our findings are beneficial even for these domains. The exploration of an additional number of shots and seeds is an interesting direction that can be explored in the future.

Sixth, we do not know if any of the 6 datasets used in this study have been used for training the LLMs we used for data collection and if this had any effect on our results and findings. As such, we do not know how much would be the comparison of established and newer LLM augmentation methods different on new, unpublished datasets. This limitation is part of the recognized possible "LLM validation crisis", as described by (Li and Flanigan, 2023).

Seventh, we used only one feature representation model for the sample selection strategies that required similarity or dissimilarity of samples, and the usage of different feature representation models could alter the performance of these sample selection strategies.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Shengnan An, Bo Zhou, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Weizhu Chen, and Jian-Guang Lou. 2023. Skill-based few-shot selection for in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13472–13492, Singapore. Association for Computational Linguistics.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024a. Effects of diversity incentives on sample diversity and downstream model performance in LLM-based text augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13148–13171, Bangkok, Thailand. Association for Computational Linguistics.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905, Singapore. Association for Computational Linguistics.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2024b. Llms vs established text augmentation techniques for classification: When do the benefits outweight the costs? *Preprint*, arXiv:2408.16502.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 8123–8144, Toronto, Canada. Association for Computational Linguistics.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *Preprint*, arXiv:2302.13007.

Luyang Fang, Gyeong-Geon Lee, and Xiaoming Zhai. 2023. Using gpt-4 to augment unbalanced data for automatic scoring. *Preprint*, arXiv:2310.18365.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023. Dale: Generative data augmentation for low-resource legal nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Sentosa, Singapore.

Abdullatif Köksal, Timo Schick, and Hinrich Schuetze. 2023. MEAL: Stable and active learning for few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 506–517, Singapore. Association for Computational Linguistics.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K. Kummerfeld. 2020. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8097–8106, Online. Association for Computational Linguistics.

Changmao Li and Jeffrey Flanigan. 2023. Task contamination: Language models may not be few-shot anymore. *Preprint*, arXiv:2312.16337.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Once: Boosting content-based recommendation with both open- and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 452–461, New York, NY, USA. Association for Computing Machinery.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*. Independently published.

Aytuğ Onan. 2023. Srl-aco: A text augmentation framework based on semantic role labeling and ant colony optimization. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101611.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024a. A survey on stability of learning with limited labelled data and its sensitivity to the effects of randomness. *ACM Computing Surveys*, 57(1).

Branislav Pecher, Ivan Srba, Maria Bielikova, and Joaquin Vanschoren. 2024b. Automatic combination of sample selection strategies for few-shot learning. *arXiv preprint arXiv:2402.03038*.

Branislav Pecher, Ivan Srba, Maria Bielikova, and Joaquin Vanschoren. 2024c. Automatic combination of sample selection strategies for few-shot learning. *Preprint*, arXiv:2402.03038.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore. Association for Computational Linguistics.

Frédéric Piedboeuf and Philippe Langlais. 2024. On evaluation protocols for data augmentation in a limited data scenario. *Preprint*, arXiv:2402.14895.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Gaurav Sahu and Issam H. Laradji. 2024. Mixsumm: Topic-based data augmentation using llms for low-resource extractive text summarization. *Preprint*, arXiv:2407.07341.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In

Proceedings of the 4th Workshop on NLP for Conversational AI, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2018. An empirical study of example forgetting during deep neural network learning. In *In International Conference on Learning Representations*.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *Preprint*, arXiv:2304.14334.

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. 2025. Diversity-oriented data augmentation with large language models. *Preprint*, arXiv:2502.11671.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *Preprint*, arXiv:2402.14568.

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.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Ethical considerations

Based on a thorough ethical assessment performed on the basis of intra-institutional ethical guidelines and checklists tailored to the use of data and algorithms, we see no ethical concerns pertaining directly to the conduct of this research. Although the production of new data through LLMs bears several risks, such as the introduction of biases, the small size of the produced dataset, sufficient for experimentation, is, at the same time, insufficient for any major machine learning endeavours where such biases could be transferred.

We follow the license terms for all the models and datasets we used (such as the one required for the use of the Llama-3.1 model) – all models and datasets allow their use as part of research.

B Model fine-tuning details

We selected the best hyper-parameters after using a hyper-parameter search. We used the same batch size across all datasets using 64 batch size, used *2e-5* learning rate, dropout 0.2, maximum number of tokens (512) trimmed and padded, and 50 number of epochs. We used AdamW optimizer in all cases. We also removed outliers of a model fine-tuning's results in some cases where the model's results were particularly unstable to account for the possible instability during training.

C Dataset details

As we did not use all of the dataset labels and samples in each of the datasets, we list our setup here. All used datasets are in English language. We either aggregated or relabelled the labels we used in datasets to ensure that datasets from all tasks of sentiment analysis, news classification, paraphrase detection, and question topic classification had the same labels. This made the out-of-distribution evaluation much easier.

We used all the labels for the *TweetEval* dataset, and for the *Yelp* dataset, we aggregated and relabelled the *one star* and *two stars* labels as *negative*, the *three stars* as *neutral* and the *four stars* and *five stars* labels as *positive*.

We used all the labels of the *AG News* dataset and for the *News Topic* dataset we aggregated and relabelled the *WORLD NEWS, POLITICS* as *U.S. NEWS* as *World*, *SCIENCE, TECH* as Science and Technology and additionally also used samples with labels *Sports* and *Business*.

For the *Yahoo* dataset, we used labels *Society & Culture, Science & Mathematics, Health, Education & Reference, Sports, Business & Finance*. We used only some labels of the Trec dataset and mapped them to the *Yahoo* dataset labels in the following way by aggregation and relabelling: on the *Society & Culture* label we mapped the *HUM:gr, HUM:ind, NUM:date, HUM:desc, ENTY:religion* labels, on the *Science & Mathematics* label we mapped the *ENTY:animal, NUM:volsize, ENTY:plant, NUM:temp* labels, on the *Health* label we mapped the *ENTY:body, ENTY:dismed* labels, on the *Education & Reference* label we mapped the *ABBR:abb, DESC:def, DESC:desc* labels, on the *Sports* label we mapped the *ENTY:sport* label and on the *Business & Finance* label we mapped the *ENTY:cremat* label.

Finally, we used all the labels in the *MNLI* dataset and the *QQP* dataset.

For the out-of-distribution split of the QQP dataset, we used the PAWS (Zhang et al., 2019) dataset, more specifically from the *labelled_final* subset and test split.

D Prompts and parameters used for LLM-based augmentation

For all of the LLMs used during augmentation, we used the same parameters: maximum number of new tokens set to 1024, sampling enabled, with *top p* set to 1 and *temperature* set to 1. We used 4-bit quantization for faster and cheaper inference on all LLMs and used instruction-tuned versions for each of the LLMs. Specifically, we used Mistral-v0.3-7B-instruct [2], Llama-3.1-8B-Instruct [3] and Gemma-2-9B-Instruct [4]. We collected 1 response and asked the LLMs to produce 5 augmentations per seed or label of that seed.

We used different prompts for generating new samples and paraphrasing existing samples. These prompts were also varied based on the dataset used.

For paraphrasing with few-shot we used this prompt: *You will be given examples from 'task' dataset, each labelled with a specific category. Based on the examples, paraphrase a given text 5 times with the 'label' category. Output each paraphrased text in the form of a numbered list separated by new lines. The text: 'text'. Examples: examples*

For paraphrasing with zero-shot, we used this prompt: *You are given a 'task' dataset. Paraphrase a given text 5 times with the 'label' category. Output each generated text in the form of a numbered list separated by new lines. The text: 'text'*

For few-shot paraphrasing of the question topic classification datasets we used this prompt: *You will be given examples of questions from 'task' dataset, each labelled with a specific topic. Based on the examples of questions, paraphrase a given question 5 times with the 'label' topic. Output each paraphrased question in the form of a numbered list separated by new lines. The question: 'text' Examples: examples*

For paraphrasing with zero-shot of the question topic classification datasets, we used this prompt:

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[4] https://huggingface.co/google/gemma-2-9b-it

You are given a 'task' dataset. Paraphrase a given question 5 times with the 'label' category. Output each generated question in the form of a numbered list separated by new lines. The question: 'text'

For few-shot paraphrasing of the MNLI dataset, we used this prompt: *You will be given a premise and hypothesis pair together with their label from a Natural Language Inference dataset. Based on the examples, paraphrase 5 times a hypothesis that 'label' the given premise. The given premise: 'premise'. Output each paraphrased hypothesis in the form of a numbered list separated by new lines. The hypothesis: 'text' Examples: examples*

For paraphrasing with zero-shot of the MNLI dataset, we used this prompt: *You will be given a premise from a Natural Language Inference dataset. Paraphrase 5 times a hypothesis that 'label' the given premise. The given premise: 'premise'. Output each paraphrased hypothesis in the form of a numbered list separated by new lines. The hypothesis: 'text'*

For few-shot paraphrasing of the QQP dataset, we used this prompt: *You will be given a question from a Paraphrase Detection dataset. Based on the examples, paraphrase 5 times a question. Output each paraphrased question in the form of a numbered list separated by new lines. The question: 'text' Examples: examples*

For paraphrasing with zero-shot of the QQP dataset, we used this prompt: *You will be given a question from a Paraphrase Detection dataset. Output each paraphrased question in the form of a numbered list separated by new lines. The question: 'text'*

For generating new samples with few-shot we used this prompt: *You will be given examples from 'task' dataset, each labelled with a specific category. Based on the examples, generate 5 new texts that fit the 'label' category. Output each generated question in the form of a numbered list separated by new lines. Examples: examples*

For generating new samples with zero-shot, we used this prompt: *You are given a 'task' dataset. Generate 5 new texts that fit the 'label' category. Output each generated question in the form of a numbered list separated by new lines.*

For few-shot generating new samples of the question topic classification datasets, we used this prompt: *You will be given examples of questions from the 'task' dataset, each labeled with a specific topic. Based on the examples of questions, generate 5 new questions that fit the 'label' topic. Output each generated question in the form of a numbered list separated by new lines. Examples: examples*

For generating new samples with zero-shot of the question topic classification datasets, we used this prompt: *You are given a 'task' dataset. Generate 5 new questions that fit the 'label' category. Output each generated question in the form of a numbered list separated by new lines.*

For few-shot generating new samples of the MNLI dataset, we used this prompt: *You will be given a premise with a label from a Natural Language Inference dataset. Based on the examples, generate 5 new hypotheses that 'label' the given premise. The given premise: 'premise'. Output each generated hypothesis in the form of a numbered list separated by new lines. Examples: examples*

For generating new samples with zero-shot of the MNLI dataset, we used this prompt: *You will be given a premise with a label from a Natural Language Inference dataset. Generate 5 new hypotheses that 'label' the given premise. The given premise: 'premise'. Output each generated hypothesis in the form of a numbered list separated by new lines.*

For few-shot generating new samples of the QQP dataset, we used this prompt: *You will be given a question from a Paraphrase Detection dataset. Based on the examples, generate 5 new questions which are 'label' considering the question. The given question: 'question'. Output each generated question in the form of a numbered list separated by new lines. Examples: examples*

For generating new samples with zero-shot of the QQP dataset, we used this prompt: *You will be given a question from a Paraphrase Detection dataset. Generate 5 new questions which are 'label' considering the question. The given question: 'question'. Output each generated question in the form of a numbered list separated by new lines.*

E Additional Results and Visualisations for Sample Selection Strategies and Their Effect on Model Performance

We provide the comparison of all sample selection strategies between each other without the baselines in Table 2. Additionally, we also provide boxplot visualization for the aggregated performance of all LLMs and random seeds in F1-Macro for models trained on various sample selection strategies together with the baselines of either random few-

shot or zero-shot for both in-distribution and out-of-distribution data in Figures 3 and 4.

F Effects of Composition of Examples and Augmentation Techniques on Model Performance

As our study had multiple parameters mentioned in Section 3, we additionally also report results for two different parameters used: *composition of examples* based on labels (using only examples from the label under augmentation or using examples from every label in the dataset) and augmentation techniques (using either *paraphrasing* of existing samples or *generation* of new samples). We report results for both parameters in Tables 3 and 4.

Each augmentation technique has the best effect on performance for either in-distribution or out-of-distribution as per Table 3. For out-of-distribution performance, the *generation* of new samples is the most often, while for in-distribution performance, the *paraphrasing* of existing samples works best. Exceptions to this are in the *Yelp* dataset, where *paraphrasing* of existing samples is best for out-of-distribution performance and *generation* of new samples for in-distribution performance.

The difference between *composition of examples* based on labels is much smaller than for augmentation techniques, as is shown in Table 4. While including samples from all labels in the dataset is better more often, the difference is quite small for out-of-distribution data. We noticed that for out-of-distribution performance, including samples from all labels worked best on question topic classification datasets and *TweetEval* dataset. In contrast, the other datasets worked better with only examples from the label under augmentation used. For in-distribution, only using examples from the target label generally leads to better downstream model performance.

DATASET→ Strategy↓	AGNEWS ID	AGNEWS OOD	NTOPIC ID	NTOPIC OOD	YAHOO ID	YAHOO OOD	TREC ID	TREC OOD	TEVAL ID	TEVAL OOD	YELP ID	YELP OOD	MNLI ID	MNLI OOD	QQP ID	QQP OOD	**TOTAL** ID	**TOTAL** OOD
Cos. sim	0	0	**2**	**3**	1	2	1	0	0	0	2	0	0	0	1	0	7	5
Cos. dis.	0	1	1	0	0	1	0	1	1	0	0	2	1	1	0	0	3	6
Forgetting most	2	0	**2**	2	1	**3**	0	1	1	1	0	0	0	0	1	0	7	7
Forgetting least	**3**	3	**2**	0	0	0	**3**	0	**3**	2	1	1	0	0	1	1	**13**	7
Carto. hard	1	0	0	1	**3**	2	0	0	1	0	1	2	0	0	1	2	7	7
Carto. easy+amb.	0	1	0	1	2	1	2	2	2	1	**3**	1	1	1	**3**	1	**13**	9
Carto. easy	1	0	1	0	2	0	1	1	1	0	2	0	1	1	1	1	10	3
Synth. dis.	2	**4**	1	2	0	0	2	**4**	0	**5**	0	**3**	**6**	**6**	1	**4**	12	**28**

Table 2: No. cases for each sample selection strategy without baseline strategies where each strategy performed the best for each dataset for in-distribution (ID) and out-of-distribution (OOD) data. The last *Total* column aggregated all cases for that specific strategy. The *Synthetic samples dissimilarity* strategy performs best on out-of-distribution model performance, while the *Cosine similarity* strategy performs best on in-distribution model performance.

Type of Augmentation	*Best for ID*	*Best for OD*
Generation	20 (27.78%)	**43 (59.72%)**
Paraphrasing	**52 (72.22%)**	29 (40.28%)

Table 3: No. cases where each type of augmentation performed the best for in-distribution (ID) and out-of-distribution (OD) data. The *generation* augmentation works best for out-of-distribution data, while the *paraphrasing* augmentation works best for in-distribution data.

Composition of Examples Type	*Best for ID*	*Best for OD*
Only From Label Under Aug.	**45 (62.5%)**	35 (48.61%)
From All Labels	27 (37.5%)	**37 (51.39%)**

Table 4: No. cases where each type of *composition of examples* type performed the best for in-distribution (ID) and out-of-distribution (OD) data. While including examples from all the labels in the dataset works best, the increase in no. cases is small.

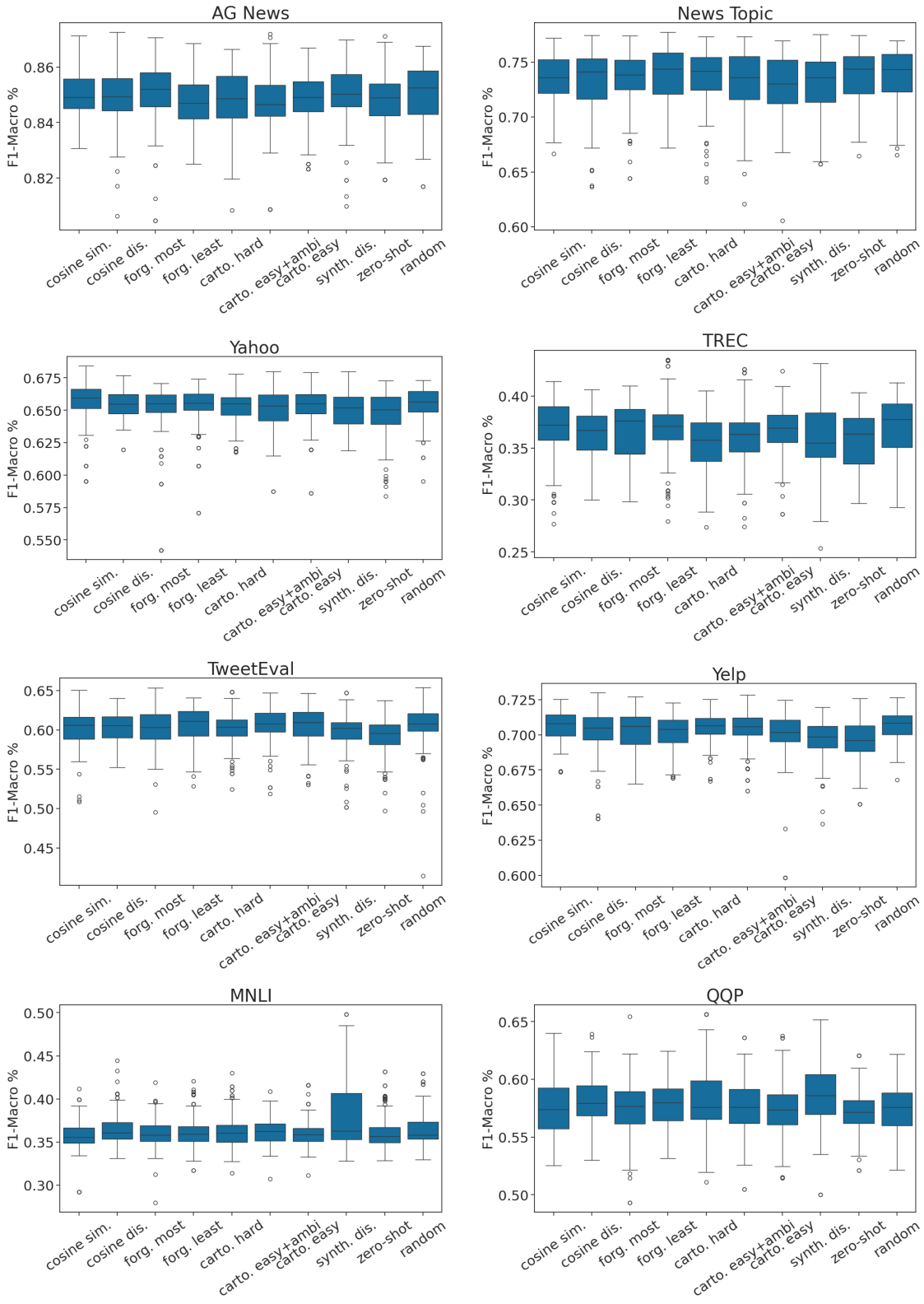

Figure 3: Aggregated performance across all LLMs and random seeds in F1-Macro for models trained on various sample selection strategies together with the baselines of either random few-shot or zero-shot on in-distribution data.

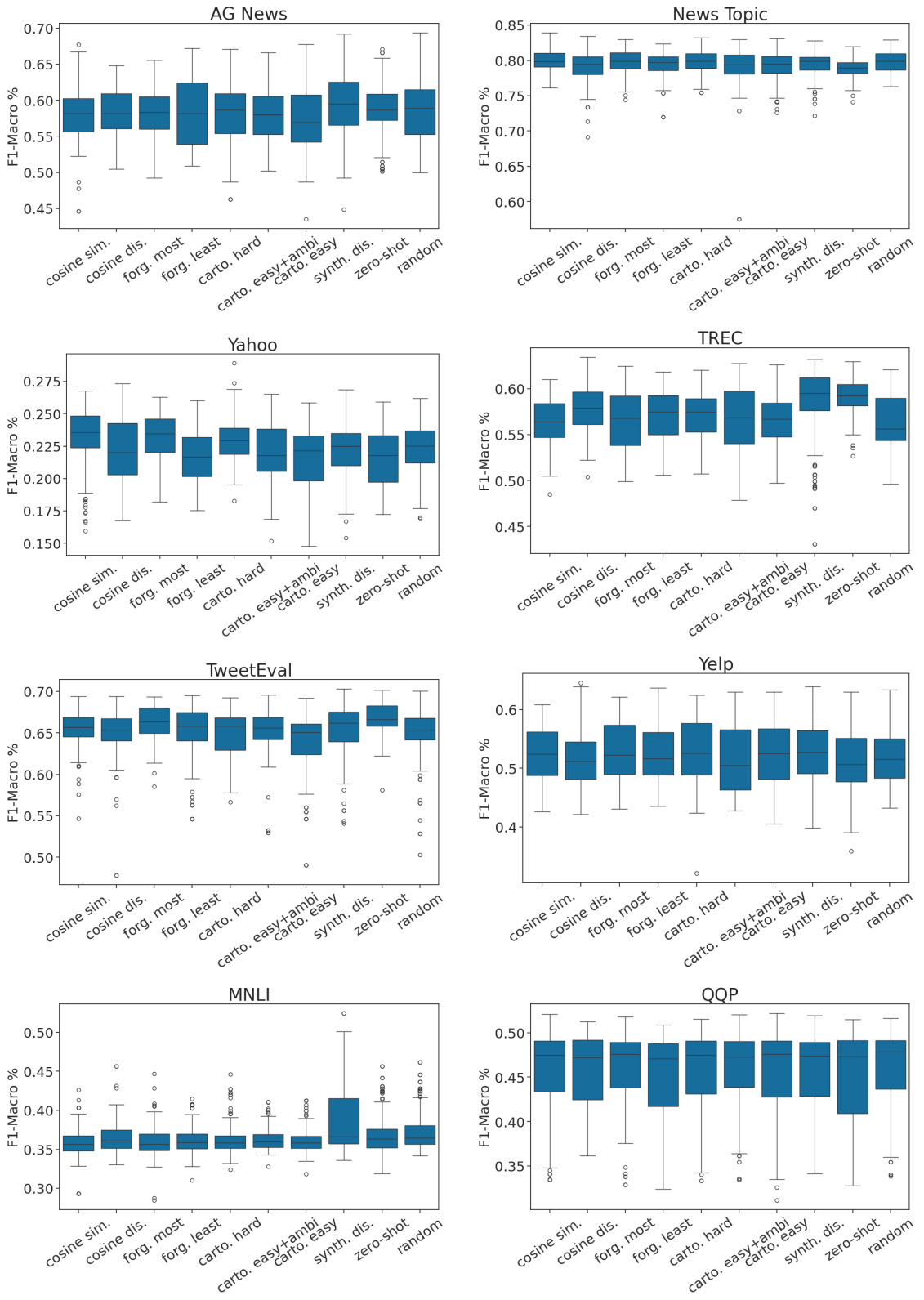

Figure 4: Aggregated performance across all LLMs and random seeds in F1-Macro for models trained on various sample selection strategies together with the baselines of either random few-shot or zero-shot on out-of-distribution data.