Boosting Zero-Shot Crosslingual Performance using LLM-Based Augmentations with Effective Data Selection

Barah Fazili*, **Ashish Sunil Agrawal***, **Preethi Jyothi**

Indian Institute of Technology Bombay, India
{barah, ashishagrawal, pjyothi}@cse.iitb.ac.in

Abstract

Large language models (LLMs) are very proficient text generators. We leverage this capability of LLMs to generate task-specific data via zero-shot prompting and promote cross-lingual transfer for low-resource target languages. Given task-specific data in a source language and a teacher model trained on this data, we propose using this teacher to label LLM generations and employ a set of simple data selection strategies that use the teacher's label probabilities. Our data selection strategies help us identify a representative subset of diverse generations that help boost zero-shot accuracies while being efficient, in comparison to using all the LLM generations (without any subset selection). We also highlight other important design choices that affect cross-lingual performance such as the use of translations of source data and what labels are best to use for the LLM generations. We observe significant performance gains across sentiment analysis and natural language inference tasks (of up to a maximum of 7.13 absolute points and 1.5 absolute points on average) across a number of target languages (Hindi, Marathi, Urdu, Swahili) and domains.[1]

1 Introduction

Multilingual pretrained models are a mainstay in modern NLP. To create highly-performant task-specific models across different languages, a commonly adopted paradigm is to finetune a multilingual pretrained model like XLM-R (Conneau et al., 2020) using task-specific labeled data. In the absence of labeled data for a target language, pretrained models finetuned on task-specific data in a source language (such as English) have been shown to facilitate zero-shot crosslingual transfer (Yu and Joty, 2021; Zheng et al., 2021; Liu et al., 2021).

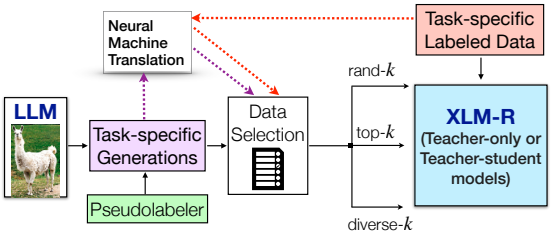

Figure 1: Overall schematic illustrating various aspects of LLM-based augmentation.

Given large language models (LLMs) and their superior generation capabilities, a natural question is whether they can be used to generate synthetic task-specific data in English. To create synthetic data in a (non-English) target language, these LLM generations could be further translated into target language data using existing machine translation systems. In this work, we examine the following central question: *How do we make best use of LLM generations to improve zero-shot*[2] *cross-lingual transfer to target languages without any labeled data?* We stress here that we are interested in the realistic setting where task-specific data in the source language might vary in domain from the target language tasks; this setting is largely absent in zero-shot evaluations in prior work.

Our overall data augmentation pipeline is illustrated in Figure 1. We use an open-source LLM such as Llama-2 (Touvron et al., 2023) and prompt it to generate task-specific text in English. For all target languages, we assume access to task-specific data in English that may not be in the same domain as the target-language tasks. When domain information is available for the target tasks, we add this information in the prompt to generate text that appears to be more in-domain.

*These authors contributed equally to this work.

[1]The code and data for this work is available at https://github.com/LLM-Based-Augmentations

[2]Here, by zero-shot we mean that we have no access to in-domain labeled data for a given target language. Automatic translations of English into the target language, derived from an NMT system as shown in Figure 1, can be used during training.

Findings of the Association for Computational Linguistics ACL 2024, pages 13406–13422

Pseudolabels for synthetic data. We examine two choices to generate pseudolabels for the LLM generations: 1. Via prompting the LLM, to appear in the output as part of the generation. 2. Via a teacher model trained on task-specific English data.

We explore two ways in which the above-mentioned pseudolabels can be used when training models with synthetic data. We train a single model on task-specific English data augmented with generations pseudolabeled via LLM prompts. We also adopt teacher-student training where the teacher is trained on task-specific English data and the student model is trained on synthetic data with soft teacher labels (i.e. label distributions). Given label noise, we find training a student model with soft teacher labels yields significantly more accurate models compared to training with hard teacher labels.

Data selection. There are two main arguments in favour of data selection:

1. Efficiency: A smaller subset of the generated data can be used as augmentation, thus reducing the training cost compared to using all the generated data.

2. Accuracy: Data selection helps identify training instances that are likely to aid learning more, and hence generalize better to yield overall performance improvements on downstream tasks.

We explore different selection strategies and show that careful data selection yields stable performance improvements, unlike random selections that lead to higher variance runs and do not guarantee gains in performance. While generating synthetic data has been studied in prior work, in this work we investigate using filtering with LLM-based augmentations in a zero-shot setting.

We note here that the role of *the teacher model is critical for data selection*. The teacher model gives label probabilities for every LLM generation, that is used in our data selection techniques. (With the pseudolabels derived via LLM prompts, we do not have such confidence estimates.) The utility of the teacher is mainly in identifying a suitable subset based on label probabilities. Once we identify such a subset, either LLM labels or teacher labels can be used for the instances in the subset.

2 Methodology

2.1 Generation

Consider a scenario where we have task-specific labeled data for a high-resource source language such as English (denoted as D_{en}). Our final downstream task is in a low-resource target language for which we have no labeled data. In this work, we experiment with two classification tasks: sentiment analysis (SA) and natural language inference (NLI). We aim to achieve improved cross-lingual transfer for these two tasks to different target languages by augmenting D_{en} with (labeled) LLM generations denoted as G_{en}. This is motivated by recent work on boosting task performance via data augmentation techniques (Vu et al., 2022; He et al., 2022a; Liu et al., 2022; Whitehouse et al., 2023; De Raedt et al., 2023a).

G_{en} is generated by prompting an LLM with a compact target domain description and the intended class label to produce class-conditioned, task-specific generations in the target domain. We utilize the open-source 13b llama-2-chat-hf model (Touvron et al., 2023) for all our generations. The prompt is composed of two sub-prompts: 1) A system prompt that specifies a generic set of rules that the generator should obey, and 2) an instruction prompt that specifies more targeted instructions for generation. More details about data generation using llama-2 and the prompts for all target tasks are specified in Appendix D and Appendix E.

2.2 Pseudolabeling and Training Methods

Teacher-student Training ($\mathcal{T}$). A teacher model is trained on the source data (D_{en}) using cross-entropy loss. The teacher is used to pseudolabel the generations in G_{en}. A subset of G_{en} is chosen via various selection techniques described in Section 2.3. We will refer to this subset as D'_{en}. A student model is trained on both D_{en} and D'_{en} combined, using cross-entropy loss with the gold labels in D_{en} and a KL-divergence loss with the soft pseudolabels derived from the teacher model. Equation (1) refers to the overall loss computed, where $y_c(x)$ is the one-hot label corresponding to each $x \in D_{\text{en}}$, q_c is the student model probability for class c (with temperature 1), $p_c(x)$ is the teacher model probability for each $x \in D'_{\text{en}}$ for class $c \in C$ and q^*_c is the student model probability for class c (scaled by temperature value 1.5). This is the standard teacher-student paradigm, and we will refer to the trained student model as $\mathcal{T}$ in our

experiments.

$$\mathrm{L}_{en} = \frac{1}{|D_{\text{en}}|} \sum_{x \in D_{\text{en}}} \sum_{c \in C} -y_c(x) \log q_c(x) + \frac{1}{|C||D'_{\text{en}}|} \sum_{x \in D'_{\text{en}}} \sum_{c \in C} p_c(x) \cdot \log \left(\frac{p_c(x)}{q^*_c(x)} \right) \quad (1)$$

Rather than using English source data and English generations, we can also adopt the translate-train setting (Artetxe et al., 2020) where D_{en} and G_{en} are translated to the target language using an off-the-shelf neural machine translation system to yield D_{tg} and G_{tg}, respectively. The rest of the above-mentioned teacher-student training pipeline stays the same, except with using translated data everywhere.

Teacher-driven Training with Prompt Labels ($\mathcal{T}_{\text{pl}}$). Instead of using a teacher model to pseudolabel the generations in G_{en}/G_{tg}, we use the teacher's label probabilities for data selection (detailed in Section 2.3) after which we label the data using the labels in the LLM prompts that we use for class-conditional generation. A single model is trained using cross-entropy loss on both source data in D_{en}/D_{tg} and prompt-labeled data sampled from G_{en}/G_{tg}. The main difference from teacher-student training is the use of hard prompt labels for the sampled generations with a cross-entropy loss instead of soft pseudolabels from a teacher model with a KL-divergence loss. Here, we first utilize the teacher for data selection and subsequently use the LLM prompt labels for the generations. This model will henceforth be referred to as $\mathcal{T}_{\text{pl}}$. Similar to $\mathcal{T}$, even with $\mathcal{T}_{\text{pl}}$, we can adopt the translate-train setting and use translated source data and LLM generations.

2.3 Data Selection Strategies

Around 130K instances are generated for each target task from which a small subset is sampled using various data selection techniques described below. In all experiments, we uniformly sample across positive, negative, and neutral class labels for sentiment analysis (and entailment, contradiction, and neutral class labels for NLI) by choosing 2500 instances from the full set of instances for each class to create D'_{en}/D'_{tg}.

- **rand-k**: We select a random subset of 2500 instances from the data generated for each class in G_{en}/G_{tg}.
- **top-k**: Instances specific to each class in G_{en}/G_{tg} are sorted in descending order using the teacher model's predicted probability for that class. The top-k ($k = 2500$) instances from each class are then selected.
- **div-k**: We aim to select a diverse set of sentences from each target class using div-k. The sentences belonging to each class (based on teacher labels) are encoded using LABSE sentence embeddings (Feng et al., 2022). The embeddings for each class are then clustered using NLTK's Kmeans clustering algorithm[3]. We create 25 clusters for each class and select the top 100 instances using the probabilities assigned by the teacher model (as in top-k) per cluster to get a total of 2500 instances per class. With this simple cluster-then-topk technique, we hope to identify samples that offer good coverage and capture the diversity of samples within each class.
- **amb-k and easy-k**: We design two additional selection techniques amb-k and easy-k by drawing inspiration from prior work on data cartography (Swayamdipta et al., 2020) where data points are characterized as ambiguous, easy or hard based on the training dynamics across epochs. We first compute predicted probabilities for each class for each instance across checkpoints of the teacher model saved after each training epoch. Next, we compute the mean and standard deviation across probabilities for each instance across training epochs. For each class, instances with the top-k ($k = 2500$) mean and standard deviation values are chosen as easy-k and amb-k, respectively. High standard deviation values signify larger variability in predictions across training; these instances are characterized as ambiguous examples that the model is unsure about. High mean values signify higher confidence in predictions; these instances are characterized as easy examples that the model is confident about. This selection technique is expensive in having to maintain checkpoints for all training epochs; we evaluate this only for NLI.

[3] https://tedboy.github.io/nlps/generated/nltk.cluster.html

3 Experimental Setup

3.1 Datasets

Source data refers to labeled task-specific data in English, while target data refers to evaluation sets in the target languages for which there is no labeled data. Unless specified otherwise, we choose source data to be from a different domain compared to the target data. This is different from most prior work in zero-shot evaluations where the source data is typically chosen to be consistent in the domain to the target tasks (Whitehouse et al., 2023; Li et al., 2021; Du et al., 2021; Vu et al., 2022). We assume a more realistic setting where the source and target domains can be mismatched.

Source data. We use SST5 (Socher et al., 2013) and SNLI (Bowman et al., 2015) datasets for sentiment analysis (SA) and natural language inference (NLI), respectively. SST5 is a sentiment classification dataset featuring five distinct labels: negative, very negative, positive, very positive, and neutral, that we collapse into three labels: positive, negative and neutral to match the target tasks. Similar to (Li et al., 2021), we consider a random subset of the SNLI train set (15K training sentences, 5K per class) to simulate a low-resource setting and for quicker experimental turnaround.

Target data. Our target SA tasks include Marathi Sentiment (Pingle et al., 2023), GLUECoS Hindi-English code-switched Sentiment (Khanuja et al., 2020), and Hindi Product Reviews (Akhtar et al., 2016). For NLI, we evaluate on Hindi, Urdu, and Swahili from the XNLI (Conneau et al., 2018) corpus; these are some of the the least-represented XNLI languages. Appendix A provides more details about the source and target tasks. Appendix C shows how we generate code-mixed data for the translate-train setting of the GLUECos task.

3.2 Model and Training details

For all our experiments, we use the xlm-roberta-large model (Conneau et al., 2019) for modeling both the student and teacher. It is a 561M parameter model.[4] Our choice of XLM-R for classification tasks was motivated by recent work on cross-lingual classification (Artetxe et al., 2023) that uses only XLM-R for all its evaluations. There is also prior work (Zhang et al., 2023) that shows that compared to much larger multilingual LMs like BLOOMZ, etc., fine-tuned models of smaller scale like XLM-R are at par or superior on many cross-lingual classification tasks for low-resource languages. Both the student and teacher models are trained for 15 epochs, with a learning rate of 5e-6, AdamW as the optimizer, batch size of 32, and gradient accumulation step size of 4. The student model uses a temperature of 1.5 for the KL-divergence loss. We use the best checkpoint model for all the evaluations, where the best checkpoint is selected based on accuracy over the source dev set. Translations are obtained using IndicTrans2 (AI4Bharat et al., 2023) for all the languages except Swahili, for which we use NLLB (Team et al., 2022).

[4] https://huggingface.co/xlm-roberta-large

3.3 Baselines

Source only (SRC). Here, the model is trained on the train set of the source tasks (refer Section 3.1). No synthetic data is used to train the model. 8,544 and 15K instances are used for SA and NLI tasks, respectively.

Source+Generations (SRC+GEN). Here, the model is trained on a mixture of source and synthetic datasets. The synthetic dataset (7.5K) is sampled randomly from among the generations and is not selected via any data selection technique or with the help of a teacher resulting in 16K and 22.5K instances for SA and NLI tasks, respectively.

Generations only (GEN). Here, we train the model only on the synthetically generated data. The labels come from the prompts that we used for class conditional generation. For a fair comparison with SRC+GEN, we maintain the same total size of 16K and 22.5K instances for SA and NLI tasks, respectively

4 Results and Analysis

4.1 Main Results

We characterize the training data used for each model along two axes: source of the task-specific text and source of labels assigned to the data. GEN, SRC and SRC+GEN in Table 1 represent the baseline models as described in Section 3.3. Generated text used in all the baseline models is combined with the corresponding prompt labels used during generation. We observe that substituting a portion of generated instances with source instances (SRC+GEN) yields better performance, as

	MarSent		HinProd		XNLI Hi		XNLI Ur		XNLI Sw		Avg
Models	dev	test	dev	test	dev	test	dev	test	dev	test	
GEN	61.13	61.64	58.86	59.43	59.98	60.66	56.61	57.35	56.71	56.23	58.86
SRC	64.17	63.56	56.98	57.94	67.67	67.72	61.65	61.34	59.16	58.54	61.87
SRC+GEN	65.70	65.91	62.11	61.67	69.74	69.29	65.28	64.64	63.52	64.30	65.22
$\mathcal{T}$-top-k	65.68	**66.53**	**68.65**	**68.80**	**71.55**	71.41	64.40	63.45	63.82	62.90	**66.72**
$\mathcal{T}$-rand-k	65.42	65.29	62.11	62.68	71.05	71.12	63.01	61.86	62.11	60.97	64.56
$\mathcal{T}$-div-k	65.81	65.99	65.55	66.83	71.26	**71.65**	65.82	64.87	64.44	64.10	66.63
$\mathcal{T}_{\text{pl}}$-top-k	**66.13**	66.02	53.98	56.50	70.40	70.39	66.27	64.06	64.38	63.78	64.19
$\mathcal{T}_{\text{pl}}$-rand-k	64.25	64.34	58.06	60.71	71.05	71.12	**66.43**	**65.18**	63.88	64.11	64.91
$\mathcal{T}_{\text{pl}}$-div-k	65.77	65.45	59.66	60.61	71.31	70.88	65.91	65.01	**65.68**	**65.34**	65.56
Delta	0.43	0.62	6.54	7.13	1.81	2.36	1.15	0.54	2.16	1.04	1.50

Table 1: This table shows the translate-train accuracies. The top three rows represent the baselines. The highest accuracy is shown in bold; the second highest is underlined. Delta represents the difference between the best-performing technique and the best-performing baseline.

-	GEN	SRC	SRC+GEN	$\mathcal{T}$-top-k	$\mathcal{T}$-rand-k	$\mathcal{T}$-div-k	$\mathcal{T}_{\text{pl}}$-top-k	$\mathcal{T}_{\text{pl}}$-rand-k	$\mathcal{T}_{\text{pl}}$-div-k	Delta
GLUECoS	52.00	51.67	54.33	53.38	53.73	55.52	49.68	48.50	49.80	1.19

Table 2: Zero-shot numbers for GLUECoS sentiment analysis

anticipated, compared to using GEN alone. Further augmenting the source data with generations (SRC+GEN) boosts the SRC baseline across all evaluated tasks/languages.

The results in Table 1 are all translate-train accuracy values since they are found to largely outperform the zero-shot numbers, thus highlighting the benefits of using (machine) translations for cross-lingual evaluations (as reported in prior work (Artetxe et al., 2023)). Please refer to Appendix B for zero-shot results. Our reported numbers are averaged across models trained on two different random seeds.[5] Following the three rows of baseline numbers in Table 1, we show results using models that are trained across two different levels of supervision from the SRC baseline (acting as the teacher). $\mathcal{T}$ indicates that data selection is done using teacher pseudolabels while $\mathcal{T}_{\text{pl}}$ indicates that after data selection using teacher pseudolabels, for each instance, prompt labels are used for subsequent training (instead of retaining the teacher-assigned labels). Each of the listed models is trained on data selected using various selection strategies detailed in Section 2.3. The $\mathcal{T}$ models are trained with soft pseudolabels derived from the teacher while the baselines and $\mathcal{T}_{\text{pl}}$ models are trained with hard prompt labels. For a given strategy (top-k, rand-k, etc.), we note that the same unlabeled data subsets are used with one of the two kinds of labels (teacher soft vs. prompt hard).

Across all tasks, we observe that data selection strategies yield consistent performance improvements over the best baseline with absolute accuracy gains of up to 7% for Hindi Product SA. $\mathcal{T}_{\text{pl}}$ appears to do better overall i.e., prompt labels after teacher-based data selection; the teacher labels perform much better just on the Hindi Product task. We note here that unlike prior work that uses LLM-based augmentations for cross-lingual tasks (Whitehouse et al., 2023) with access to some target data, all our models are trained without *any access* to real target data.

We find that the delta values in Table 1 using our data selection techniques for XNLI and Marathi SA (that have a similar number of test instances) are statistically significant at $p < 0.01$ using the Wilcoxon signed rank test. Since the Hindi product review task has a significantly smaller number of test instances, we treat it separately across different random seeds and find that top-k data selection results in a statistically significant improvement (compared to SRC+GEN) at $p < 0.05$ using the Wilcoxon signed rank test.

Other than the five target sets in Table 1, in Table 2 we also evaluate on a code-switched Hindi-English sentiment analysis task which is yet another challenging low-resource domain. Unlike

[5]For the Hindi SA task, due to the considerably smaller size of the evaluation sets, we trained the models using six different seeds to obtain more reliable evaluations.

	XNLI Hi		XNLI Ur		XNLI Sw	
	dev	test	dev	test	dev	test
SRC+GEN	69.74	69.29	65.28	64.64	63.52	64.3
$\mathcal{T}$-amb-k	69.14	70.03	64.00	62.49	63.98	63.02
$\mathcal{T}$-easy-k	69.36	69.75	65.47	64.85	63.17	62.76
$\mathcal{T}_{pl}$-amb-k	**72.05**	**71.36**	**66.91**	**66.12**	**65.66**	**64.98**
$\mathcal{T}_{pl}$-easy-k	70.62	70.59	63.82	62.88	64.30	63.49
Delta	2.31	2.07	1.63	1.48	2.14	0.68

Table 3: Translate-train accuracies using amb-k and easy-k selection (see section 2.3). Delta is (best score - SRC+GEN) score.

	Marsentiment		XNLI Hi	
	$\mathcal{T}$-top-k		$\mathcal{T}$-top-k	
	dev	test	dev	test
CE	45.75	44.43	40.52	40.09
KLD	**65.68**	**66.53**	**71.55**	**71.41**
Delta	19.93	22.10	31.03	31.32

Table 4: Compare training of $\mathcal{T}$-top-k model with teacher-soft vs teacher-hard labels. Delta represents the difference between the accuracy (translate-train) for the student trained via soft labels and the student trained using hard labels.

Table 1 where translate-train was more effective, we show zero-shot scores since the presence of English words in code-switched En-Hi text is found to benefit more from SRC+GEN (teacher) trained on original generations in English. More details about the code-mixed text generation and the corresponding Translate-train numbers are provided in Appendix C.

4.2 Experimental Analysis

Ambiguous/Easy Data Selection. Table 3 shows XNLI results of student models/prompt-based models trained on data selected using amb-k and easy-k selection techniques. Augmenting the source data with prompt-labeled ambiguous instances benefits the model the most. Ambiguous instances are ones that the model is most uncertain about and are likely to help the model generalize well. This is consistent with observations about ambiguous instances in prior work (Swayamdipta et al., 2020; Liu et al., 2022).

Soft Labels vs. Hard Labels. Table 4 shows the translate-train accuracies of a student model ($\mathcal{T}$) trained using teacher hard pseudo labels and soft pseudo labels. CE implies training using teacher hard labels and cross-entropy loss, KLD implies training using teacher soft labels and KL-divergence loss. Large delta values highlight the significance of using soft teacher labels instead of hard labels. If the teacher labels are noisy, soft labels help the student model generalise better to unseen data.

Effect of Varying Sizes of Augmented Data. To study the effect of augmented data size on cross-lingual transfer, we experiment with div-k selection ($\mathcal{T}$ model) and SRC+GEN model for the Marathi SA task. Models are trained in the translate-train setting over varying amounts of augmented data. Table 5 shows that increasing k leads to a decrease in accuracy. The consistent downward trend in the div-k selection technique underscores the importance of data selection; the best accuracies were obtained using div-k with $k = 7500$. Also, augmenting synthetic data also results in increased training time. Determining the optimal augmentation size for each target task is left as future work.

Augmenting with Target Train Data. To explore whether LLM generations boost performance even in the presence of source data that matches in domain to the target task (henceforth referred to as target training data), we train teacher models on a subset of 15K sentences from the XNLI train set. (Note that the numbers in Table 1 were obtained using SNLI as the source data.) As expected, we see significant improvements in teacher accuracies in Table 6 when using target train data. Student models trained on generations pseudolabeled with this superior teacher further boost accuracies; the best results are obtained using a combination of

	Marsentiment			
	$\mathcal{T}$-div-k		SRC+GEN	
	dev	test	dev	test
$k = 2500$	65.81	65.99	**65.38**	65.42
$k = 7500$	**66.32**	**66.57**	64.31	64.44
$k = 12500$	65.28	65.17	63.57	64.39
$k = 17500$	64.73	64.53	65.14	**65.57**
$k = 22500$	64.64	64.22	64.08	64.79
Delta	1.68	2.35	1.30	0.78

Table 5: Translate-train accuracy analysis by training $\mathcal{T}$-div-k and SRC+GEN models with different sizes of augmented data (k is the examples augmented per class). Delta represents the difference between the best accuracy and accuracy for $k = 22500$.

	XNLI Hi		XNLI Ur		XNLI Sw	
	dev	test	dev	test	dev	test
SRC	79.60	78.10	71.81	70.00	73.37	72.28
SRC+GEN	79.32	78.47	72.13	69.96	73.34	72.33
$\mathcal{T}$-top-k	79.78	77.82	72.31	70.00	**74.24**	**72.41**
$\mathcal{T}$-div-k	**80.44**	**78.54**	**72.87**	**70.89**	74.12	71.70
Delta	0.84	0.44	1.06	0.89	0.89	0.13

Table 6: Using in-domain target train (taking a random subset from MNLI) for training the Teacher and also as the source data (translate-train accuracies).

	$\mathcal{T}$-top-k	
	dev	test
2500 pos, 2500 neg, 2500 neu	**65.68**	**66.53**
3000 pos, 3000 neg, 1500 neu	64.28	65.40
3500 pos, 3500 neg, 500 neu	65.13	65.51
3750 pos, 3750 neg, 0 neu	64.98	65.08

Table 7: Probing the effect of class imbalance on Marathi sentiment task. pos, neg, and neu indicate positive, negative, and neutral classes. The best numbers (translate-train) in a column are highlighted.

translate-train and div-k selection. Also, the accuracies obtained using SRC+GEN models are inferior to the best scores obtained using $\mathcal{T}$ models, again highlighting the importance of data selection.

Generations Uniform across Classes. By default, we create class-balanced augmentations by sampling 2500 instances from each class based on teacher labels. To analyse the effect of class imbalance on downstream evaluation, we augment the sentiment source data with class-imbalanced augmented data sets for the Marathi SA task. The total size of the augmented data remains constant, while the class distribution is altered by eliminating neutral sentences from 2500 to 0, thereby transitioning towards sentiment-rich augmentations. We see in Table 7 that students trained on data that is uniformly distributed across classes along with the top-k selection strategy exhibits superior performance, compared to those trained on a subset of the generated data with imbalanced class proportions. This suggests that employing a class-balanced augmentation is an important consideration.

Quality of prompt labels. To evaluate whether the LLM prompt labels are truly reflected in the filtered generated text or not, we ran a human evaluation on a set of 100 generations each for sentiment analysis and NLI. These instances were randomly

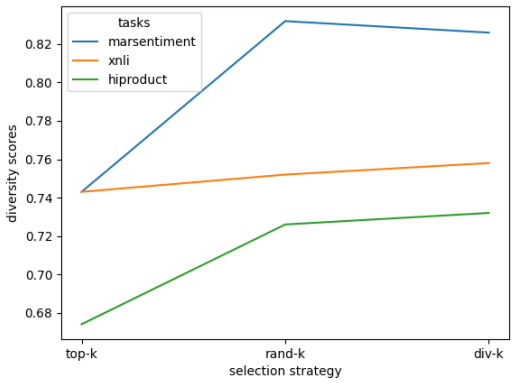

Figure 2: Diversity scores of augmented data for different data selection strategies and different tasks.

selected generations from among a set of pseudolabels predicted with high probability by the teacher model for the respective tasks. The average accuracy of label alignment between annotator-provided labels and prompt-derived labels was found to be 87.88% and 71.72%, with Cohen's kappa coefficients of 0.752 and 0.749, respectively for the two tasks. This suggests that the prompt labels in the subset obtained after teacher-based filtering are of fairly high quality. Please refer to Appendix G for more details.

Analyzing Diversity. We introduce a simple metric that we call a *"diversity score"* to capture the dissimilarity in text embeddings across sentences in a dataset. This is computed by encoding each instance using LABSE (Feng et al., 2022), taking the average of the cosine distance of the LABSE embedding with every other instance in the data sample and finally taking an average of these mean distances across all data samples. To check if the data selected using the div-k selection technique is indeed diverse, we compute the diversity score for each task and data selection strategy. Figure 2 shows the trend of diversity scores. It is clear that the diversity of the 7500 sentences selected using div-k technique is greater than the diversity of the sentences selected via top-k and rand-k across all tasks.

Cross-Domain Analysis. Recall that the prompts to the LLM also contained domain information of the target task. To evaluate the impact of domain-specificity of the generations on zero-shot performance, we create two cross-domain datasets in the medical and law domain (unrelated to the target task domains). Table 8 shows zero-shot results of

	MarSent		XNLI Hi	
	dev	test	dev	test
SRC+GEN-top-k	65.38	65.42	70.10	70.20
$\mathcal{T}$-top-k (in-domain)	<u>66.34</u>	**66.54**	70.82	69.81
$\mathcal{T}$-top-k law	**66.45**	<u>66.47</u>	**71.43**	**70.08**
$\mathcal{T}$-top-k medical	66.00	64.83	<u>71.26</u>	<u>69.97</u>

Table 8: Cross-domain experiments for Marathi sentiment and XNLI-Hi task with top-k selection. Best numbers are highlighted and the second-best are underlined.

	Pos score	**Neg score**	**Overall score**
Medical	0.037	0.035	0.073
Law	0.034	0.027	0.061
Marathi In-d	0.026	0.020	0.046

Table 9: Sentiment richness of Marathi sentiment in-domain and out-of-domain datasets; Pos, Neg indicates positive and negative.

the student models trained on the in-domain and out-of-domain (law, medical) augmented datasets for the Marathi sentiment and XNLI Hi tasks. We observe that the student model trained on out-of-domain datasets perform comparably to the student models trained on in-domain data. This shows that models can effectively learn task information from augmented data even if it comes from domains that differ from the target task.

To further disentangle the roles of task and domain, we made an attempt to capture sentiment richness present in the in-domain vs out-of-domain data for the Marathi SA task. We use a sentiment lexicon[6] consisting of a positive and negative sentiment score for each word along with its POS tag, word sense, etc. We simply add the corresponding scores for each word in the corpus found in the lexicon (along with matching the POS tag) and normalize the sums by the word count of the corpus. Table 9 shows that the computed sentiment scores for the generated out-of-domain datasets are much better than the in-domain dataset for Marathi SA. Here, overall score is calculated by summing both the positive and negative scores for each word. Hence, it is plausible that the sentiment richness in the generated out-of-domain datasets compensates for the domain mismatch, thus yielding comparable or slightly better results in Table 8. (Tables 28, 29 in Appendix E show examples of generations from different domains.)

[6]We used the SentiWordNet 3.0 sentiment lexicon available at: https://github.com/aesuli/SentiWordNet?tab=readme-ov-file).

5 Related Work

Our work is closely related to Whitehouse et al. that studies generations from various open-source and commercial LLMs for cross-lingual performance over reasoning tasks. However, they rely on instances from the target sets as few-shots for generations and do regular finetuning over labels derived from the class-conditional prompts.

Furthermore, our work is grounded in ideas inspired by He et al., incorporating self-training on unlabeled synthetic text produced by Language Models. However, He et al. fine-tuned generators using target data, and their experiments were limited to GLUE tasks (in English). In contrast, our focus is on multilingual models, aiming for cross-lingual transfer from source data in a high-resource language across arbitrary domains in different task languages. We achieve this by self-training on zero-shot generations from LLMs without utilizing any target data during generation or training.

We draw inspiration from (Swayamdipta et al., 2020) to design data selection techniques based on the principle of dataset cartography. (Liu et al., 2022) use dataset cartography on a large NLI dataset (MNLI) to choose instances with complex reasoning patterns, and instructs GPT-3 to generate new examples with similar patterns. Automatically generated examples undergo filtering, and ultimately, human crowdworkers review, revise, and label them. In a similar vein, Khanuja et al. present language-agnostic methods to pick specific data points to be labeled from a large, unlabelled multilingual dataset. These points are chosen either by considering their distance from the target set, the uncertainty of model predictions over them, or finding a balance between minimizing distance and maximizing model uncertainty. While WANLI (Liu et al., 2022) only explores English generation/evaluation, both depend not only on human-in-the loop annotation of unlabeled text, but also depend on existing target data for generator finetuning (Khanuja et al., 2023) or few-shot prompting (Liu et al., 2022).

De Raedt et al. propose in-place augmentation of data instances from high-resource languages for better out-of-distribution generalization by leveraging LLMs. However, these techniques are specifi-

cally applicable to single-text classification tasks. For tasks like NLI or QA, which involve multiple components in each instance, automatically making such edits or augmentations while preserving the intended relationships between components (and affecting labels) is not straightforward.

In contrast to (Li and Callison-Burch, 2023), and (Riabi et al., 2021), we make use of an open-source LLM to generate task and domain-specific synthetic data. We also explore the more realistic setting of having no access to source data that matches the target domain; this is not explored in the above two works. Similar to our work, synthetic data generation has been explored in (Agrawal et al., 2023; Gekhman et al., 2023). However, both works do not use any form of data selection for the synthetically generated data. (Gekhman et al., 2023) do not show evaluations on low-resource languages; their multilingual experiments are limited to high-resource languages such as English, French, Spanish and German. (Agrawal et al., 2023) show experiments in the few-shot setting, while we operate in the zero-shot setting.

6 Conclusion

In this work, we focus on the broader problem of boosting zero-shot cross-lingual transfer using LLM-based augmentations. We highlight the importance of using data selection strategies to select smaller subsets that result in more efficient training and improved performance on downstream target language tasks. We also compare and contrast the utility of pseudolabeling generations using labels from LLM prompts versus using a teacher model to label the generations. One of the main takeaways is that LLM generations, in conjunction with our data selection strategies, can help improve cross-lingual transfer regardless of whether task-specific source data matches the domain of the target tasks or not.

Acknowledgements

The authors thank the anonymous reviewers for their constructive feedback and engagement during the rebuttal that helped improve the quality of the draft. The last author would like to gratefully acknowledge a faculty grant from Google Research India supporting her research on multilingual models.

Limitations

1. The generations from LLMs are sensitive to the prompts used. Although we share our custom prompts, the quality of the generated content is heavily reliant on the particular domain and task for which the data is generated creating some non-determinism.

2. Because of budget constraints, our investigations were constrained to an open-source LLM (LLAMA-2). It is possible that higher-capacity commercial LLMs could yield better performance.

3. We explore many data selection techniques but a clear winner across all tasks/settings has not emerged. Although ambiguous selection gives best scores for XNLI, more target domains and languages should be included to study the most effective filtering techniques in general.

4. We have only experimented with generating data for classification tasks; generating data for more structured tasks like QA or commonsense reasoning tasks could pose challenges.

5. For the translate-train models, one assumes access to MT models for the target language which may not always be available.

References

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. Qameleon: Multilingual qa with only 5 examples.

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709, Portorož, Slovenia. European Language Resources Association (ELRA).

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Maarten De Raedt, Semere Kiros Bitew, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023a. Zero-shot cross-lingual sentiment classification under distribution shift: an exploratory study. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 50–66, Singapore. Association for Computational Linguistics.

Maarten De Raedt, Semere Kiros Bitew, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023b. Zero-shot cross-lingual sentiment classification under distribution shift: an exploratory study. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 50–66, Singapore. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022a. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022b. Generate, annotate, and learn: Nlp with synthetic text.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos : An evaluation benchmark for code-switched NLP. *CoRR*, abs/2004.12376.

Simran Khanuja, Srinivas Gowriraj, Lucio Dery, and Graham Neubig. 2023. Demux: Data-efficient multilingual learning.

Bryan Li and Chris Callison-Burch. 2023. PAXQA: Generating cross-lingual question answering examples at training scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 439–454, Singapore. Association for Computational Linguistics.

Shiyang Li, Semih Yavuz, Wenhu Chen, and Xifeng Yan. 2021. Task-adaptive pre-training and self-training are complementary for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1006–1015, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational*

Linguistics: EMNLP 2022, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71, Online. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Sneha Mondal, Ritika, Shreya Pathak, Preethi Jyothi, and Aravindan Raghuveer. 2022. CoCoa: An encoder-decoder model for controllable code-switched generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3cube-mahasent-md: A multi-domain marathi sentiment analysis dataset and transformer models.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. 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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Tu Vu, Minh-Thang Luong, Quoc V. Le, Grady Simon, and Mohit Iyyer. 2022. Strata: Self-training with task augmentation for better few-shot learning.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced cross-lingual performance.

Tao Yu and Shafiq Joty. 2021. Effective fine-tuning methods for cross-lingual adaptation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8492–8501, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023*

Task	#Train	#Dev	#Test
SST5	8,544	1,101	2,210
SNLI	15,000	9,842	9,824

Table 10: Source task details

Task	#Dev	#Test	Language	Source Dataset
Marathi Sentiment	6,000	6,750	Marathi	SST5
IITP Product Review	523	523	Hindi	SST5
XNLI	2,490	5,010	Hindi	SNLI
XNLI	2,490	5,010	Urdu	SNLI
XNLI	2,490	5,010	Swahili	SNLI
GLUECos Sentiment	1,260	-	Hinglish	SST5

Table 11: Target task details

Conference on Empirical Methods in Natural Language Processing, pages 12567–12582, Singapore. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. 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 3403–3417, Online. Association for Computational Linguistics.

A Dataset Description

Table 10 shows the details of the source tasks. SST5 is used as a source task for all the sentiment tasks, and SNLI as a source task for the XNLI tasks. Table 11 shows the details of the different target tasks. We treat both the dev and test sets of these target tasks as test sets and do not use their train sets for model training. The language "Hinglish" refers to the Hindi-English code-mixed text, the data is present in a mixed script (both Romanized and Devanagari) form. Table 12 shows the domain information present in the different target tasks. The domain list is not exhaustive, we made use of the domains that were easier to represent in a prompt.

B Zero-shot evaluations

Tables 13, 15, 16, 17 shows the zero-shot results of the baselines and our techniques. In table 13, except for a few languages, our techniques beat the baselines by a reasonable amount. Table 14 shows the translate-train numbers of the GLUECos task. The evaluation set of GLUECos is in romanized Hindi-English (code-switched text), we suspect this to be the reason for the negative delta.

MarSentiment	Mixture of political tweets, sitcom subtitles, generic tweets and movie reviews
Hindi Product	Reviews about travel, movies, and various electronic gadgets
XNLI	Travel, fiction, government domains
GLUECos	Generic tweet domain

Table 12: Target Task domains, the list is not exhaustive, we picked up domains which could be represented well in a prompt

C Code-mixed Text generation

To generate code-mixed text in mixed-script format for the GLUECos sentiment analysis task, we trained a mt0-xl (Muennighoff et al., 2023) model. The pre-trained model weights were downloaded from the HuggingFace repo[7]. The training is done using the huggingface's Trainer API. We made use of Lora (Hu et al., 2021) PEFT technique (Mangrulkar et al., 2022) to train the model, with the hyperparameters: alpha=32, dropout=0.05, r=16, and Lora matrices being applied to query, key, and value attention matrices. The model is trained for 15 epochs, train batch size of 16, gradient accumulation steps of 4, learning rate of 2e-4, max grad norm of 0.3, and warmup ratio of 0.03. The best model checkpoint was selected using the evaluation loss. The data for training the model was obtained from (Mondal et al., 2022). Table 18 shows few examples. Before generating the translate-train code-mixed text, we first translate the SST5 English train set to Hindi, which we then feed into our model to get the Hinglish mixed-script outputs.

D LLAMA-2 Generation Details

We use LLAMA-2 13b-chat-hf model for all our generations, as it is a recent state-of-the-art and open-source model. We use a 4-bit quantized version of the model owing to memory constraints. For quantization, we make use of the Bitsandbytes library (Dettmers et al., 2022). We use nucleus

[7]https://huggingface.co/bigscience/mt0-xl

	MarSent		HinProd		XNLI Hi		XNLI Ur		XNLI Sw	
selection strategy	dev	test	dev	test	dev	test	dev	test	dev	test
GEN	62.28	62.66	60.96	61.38	62.89	62.06	57.21	56.59	56.02	56.17
SRC	63.70	62.96	61.95	63.48	65.54	64.61	57.15	55.43	57.63	55.45
SRC+GEN	65.38	65.42	65.14	63.16	70.10	70.20	**63.92**	**63.81**	63.34	**63.64**
$\mathcal{T}$-top-k	66.34	66.54	**66.96**	**69.00**	70.82	69.81	62.98	62.02	63.74	62.94
$\mathcal{T}$-rand-k	66.22	66.06	66.03	67.56	68.80	67.68	59.78	58.21	61.67	59.49
$\mathcal{T}$-div-k	66.38	66.12	66.25	68.04	70.38	69.65	62.83	61.03	62.69	61.51
$\mathcal{T}_{pl}$-top-k	**67.83**	**68.01**	58.29	58.67	70.02	69.86	63.33	61.73	62.11	61.18
$\mathcal{T}_{pl}$-rand-k	65.16	65.05	60.77	59.27	**71.49**	**71.19**	63.78	63.29	63.24	62.49
$\mathcal{T}_{pl}$-div-k	63.13	63.18	60.77	59.85	70.00	69.89	62.33	62.27	**63.88**	61.86
Delta	2.45	2.59	1.82	5.84	1.39	0.99	-0.14	-0.52	0.54	-0.70

Table 13: This table shows the zero-shot accuracies. The top three rows represent the baselines. The highest accuracy is shown in bold; the second highest is underlined. Delta represents the difference between the best-performing technique and the best-performing baseline

-	GEN	SRC	SRC+GEN	$\mathcal{T}$-top-k	$\mathcal{T}$-rand-k	$\mathcal{T}$-div-k	$\mathcal{T}_{pl}$-top-k	$\mathcal{T}_{pl}$-rand-k	$\mathcal{T}_{pl}$-div-k	Delta
dev	51.62	53.02	55.05	52.42	50.44	52.62	48.18	49.09	48.26	-2.43

Table 14: Translate-train numbers for GLUECoS Sentiment Analysis

	XNLI Hi		XNLI Ur		XNLI Sw	
	dev	test	dev	test	dev	test
SRC+GEN	70.10	70.20	**63.92**	**63.81**	63.34	**63.64**
$\mathcal{T}$-amb-k	68.68	67.47	58.28	57.19	59.32	56.55
$\mathcal{T}$-easy-k	66.43	65.20	57.95	56.38	59.28	57.05
$\mathcal{T}_{pl}$-amb-k	**72.53**	**71.67**	63.90	63.58	**63.86**	63.25
$\mathcal{T}_{pl}$-easy-k	69.68	69.21	62.67	61.22	61.59	61.49
Delta	2.43	1.47	-0.02	-0.23	0.52	-0.39

Table 15: Zero-shot accuracies using amb-k and easy-k selection (see section 2.3). Delta is (best score - SRC+GEN) score.

	Marsentiment		XNLI Hi	
	$\mathcal{T}$-top-k		$\mathcal{T}$-top-k	
	dev	test	dev	test
CE	33.72	33.65	43.70	43.28
KLD	**66.34**	**66.54**	**70.82**	**69.81**
Delta	32.62	32.89	27.12	26.53

Table 16: Compare training of $\mathcal{T}$-top-k model with teacher-soft vs teacher-hard labels. Delta represents the difference between the accuracy (zero-shot) for the student trained via soft labels and the student trained using hard labels.

	$\mathcal{T}$-top-k	
	dev	test
2500 pos, 2500 neg, 2500 neu	66.34	**66.54**
3000 pos, 3000 neg, 1500 neu	64.20	63.94
3500 pos, 3500 neg, 500 neu	65.59	65.11
3750 pos, 3750 neg, 0 neu	**66.75**	66.18

Table 17: Probing the effect of class imbalance on Marathi sentiment task. pos, neg, and neu indicate positive, negative, and neutral classes. The best numbers (zero-shot) in a column are highlighted.

sampling for all our generations with p=0.9, we avoid repeating bi-grams, and we keep a temperature between 1.5-2.5 depending on the response of the model to the input prompt. We generated approximately 2lac sentences for each task, which came down to 1.5lac sentences post-processing and cleaning. We took a random subset of 1.3 lac sentences from the total generations, to which we then applied data selection techniques.

E LLAMA-2 Prompt Details

Tables 19, 20, 21, 22, 23, 24, 25, 26, 27 show the prompts we used to generate data for different target tasks.[8] Custom prompts are employed for each

[8]We also tried generating task-specific sentences without specifying any domain information. This resulted in noisier generations compared to when we added a domain description in the prompt. This might be an artefact specific to the LLM

target class. In the case of the Hindi product, where the goal is to generate product-specific reviews, numerous neutral (class-conditioned) generations still exhibited subtle indications of positive or negative sentiments. Consequently, we generated additional neutral sentences using a slightly modified prompt to achieve a balanced distribution of classes in the augmented dataset.

For the Marathi Sentiment target task, we devise specific prompts for various domains such as political tweets, generic tweets, subtitles, and movie reviews. To create premise-hypothesis pairs for XNLI, we initially generate the premises providing domain information in the prompts. The hypothesis is then generated using the respective premise as input along with the NLI label (entailment, neutral, and contradiction). Tables 28, 29 show the generations belonging to different domains for sentiment classification and natural language inference (NLI) task.

F Computational Budget

Training a student model on an Nvidia A100 GPU with 80G of RAM took ~100 mins for 15 epochs. We utilized the same GPUs for text generation. Generating ~2 lac sentences of max-length 256 tokens using LLAMA-2 13B model with a batch size of 60 took ~20 hrs using a single GPU and ~50 GB of RAM.

G Annotator Details

Both annotators had professional competence in English. The instructions given for the two tasks are listed below:

1. Sentiment Task: Given a sentence, identify the polarity of sentiment which could be one of the three types: positive, negative and neutral.

2. NLI Task: Given two pieces of text called premise and hypothesis, mark the pair as "entailment" if premise entails the hypothesis, "contradiction" if premise contradicts the hypothesis and "neutral" if there's neither entailment nor contradiction i.e the factuality of two statements is independent from each other.

we used, and needs further investigation.

Input text	Code-mixed text
Ye baccho ko le jane layak hai.	Ye *children* ko le jane layak hai.
vinsent gailo is phraanseesee shokar mein ghar par apne saamaany bure ladake kee ajeeb bhoomika nibha raha hai.	*Vincent Galo* is *French* shokar mein ghar par apne *normal bad boy* ki *odd role* nibha raha hai.

Table 18: Examples of code-mixed generation by our trained model. The above examples are transliterated for ease of reading, the words which are translated to English by the model are emphasized.

Positive	Negative	Neutral	Neutral Add.
<s>[INST] «SYS» You are a user providing reviews on travels, movies and various electronic gadgets. Please only generate the review without any additional content before or after. «/SYS» Please generate a single review in not more than two short sentences on one of the system specified products/movies/travels indicating a positive sentiment.[/INST]	<s>[INST] «SYS» You are a user providing reviews on travels, movies and various electronic gadgets. Please only generate the review without any additional content before or after. «/SYS» Please generate a single review in not more than two short sentences on one of the system specified products/movies/travels indicating a negative sentiment.[/INST]	<s>[INST] «SYS» You are a user who talks about travels, movies and various electronic gadgets in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the description without any additional content before or after. «/SYS» Please generate a single and very short sentence on one of the system specified products/movies/travels. It should provide very general information.[/INST]	<s>[INST] «SYS» You are a user who talks about travels, movies and various electronic gadgets in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the description without any additional content before or after. «/SYS» Please generate a single and very short sentence on one of the system specified products/movies/travels. It should provide very general information, strictly do not use positive words or adjectives.[/INST]

Table 19: Prompts for HinProduct Task

Positive	Negative	Neutral
<s>[INST] «SYS» You are a user, political figure or activist who tweets a variety of thoughts and perspectives on current affairs. Please only generate the tweet without any additional content before or after. «/SYS» Please generate a single tweet in not more than two sentences indicating a positive sentiment with no hashtags, and minimal noise.[/INST]	<s>[INST] «SYS» You are a user, political figure or activist who tweets a variety of thoughts and perspectives on current affairs. Please only generate the tweet without any additional content before or after. «/SYS» Please generate a single tweet in not more than two sentences indicating a negative sentiment with no hashtags, and minimal noise.[/INST]	<s>[INST] «SYS» You are a user, political figure or activist who tweets a variety of thoughts and perspectives on current affairs in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the tweet without any additional content before or after. «/SYS» Please generate a single tweet in not more than two sentences with no hashtags, and minimal noise. It should provide very general information.[/INST]

Table 20: Prompts for MarSentiment Task, specific to political tweets

Positive	Negative	Neutral
<s>[INST] «SYS» Please generate subtitles from any situational comedy TV show. Please only generate the subtitle without any additional content before or after. «/SYS» Please only generate a single sentence taken from the subtitles indicating a positive sentiment without specifying the speaker or any other plot details.[/INST]	<s>[INST] «SYS» Please generate subtitles from any situational comedy TV show. Please only generate the subtitle without any additional content before or after. «/SYS» Please only generate a single sentence taken from the subtitles indicating a negative sentiment without specifying the speaker or any other plot details.[/INST]	<s>[INST] «SYS» Please generate subtitles from any situational comedy TV show in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the subtitle without any additional content before or after. «/SYS» Please only generate a single sentence taken from the subtitles without specifying the speaker or any other plot details. It should provide very general information.[/INST]

Table 21: Prompts for MarSentiment Task, specific to subtitles

Positive	Negative	Neutral
<s>[INST] «SYS» You are a user who tweets on a variety of domains. Please only generate the tweet without any additional content before or after. «/SYS» Please generate a single tweet in not more than two sentences indicating a positive sentiment with no hashtags, and minimal noise.[/INST]	<s>[INST] «SYS» You are a user who tweets on a variety of domains. Please only generate the tweet without any additional content before or after. «/SYS» Please generate a single tweet in not more than two sentences indicating a negative sentiment with no hashtags, and minimal noise.[/INST]	<s>[INST] «SYS» You are a user who tweets on a variety of domains in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the tweet without any additional content before or after. «/SYS» Please generate a single tweet in not more than two sentences with no hashtags, and minimal noise. It should provide very general information.[/INST]

Table 22: Prompts for MarSentiment Task, specific to generic tweets

Positive	Negative	Neutral
<s>[INST] «SYS» You are a user who provides reviews on a variety of Indian movies. Please only generate the review without any additional content before or after. «/SYS» Please generate a single review in not more than two sentences indicating a positive sentiment with minimal noise.[/INST]	<s>[INST] «SYS» You are a user who provides reviews on a variety of Indian movies. Please only generate the review without any additional content before or after. «/SYS» Please generate a single very short review in not more than two sentences indicating a negative sentiment with minimal noise.[/INST]	<s>[INST] «SYS» You are a user who provides reviews on a variety of Indian movies in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the review without any additional content before or after. «/SYS» Please generate a single, short review in not more than two sentences with minimal noise. It should provide very general information.[/INST]

Table 23: Prompts for MarSentiment Task, specific to movie reviews

Travel	Government	Fiction
<s>[INST] «SYS»You are a user who talks about other people's traveling experiences. Please only generate the traveling experience in a single sentence without any additional content before or after. «/SYS»Please generate a single and short sentence belonging to the travel domain.[/INST]	<s>[INST] «SYS»Your job is to generate a diverse sentence in the domain provided by the user. Please only generate the sentence without any additional content before or after. «/SYS»Please generate a single and short sentence belonging to the government domain. [/INST]	<s>[INST] «SYS»Your job is to generate a diverse sentence in the domain provided by the user. Please only generate the sentence without any additional content before or after. «/SYS»Please generate a single and short sentence belonging to the fiction domain. [/INST]

Table 24: Prompts for XNLI premises.

Entailment	Neutral	Contradiction
<s>[INST] «SYS»Please generate a single sentence that is implied from the sentence provided by the user. The sentence generated could encompass either a portion or the entirety of the information contained in the given sentence. Please ensure the generations are grammatically correct. Please only share the generation without any additional content before or after.«/SYS»Sentence: [/INST]	<s>[INST] «SYS»Please generate a single sentence related to the user provided sentence that is neither entailed nor contradicts the user provided sentence. Please ensure the generations are grammatically correct. Please only share the generation without any additional content before or after.«/SYS»Sentence: [/INST]	<s>[INST] «SYS»Please generate a single sentence that logically contradicts the information provided in the sentence given by the user. Please ensure the generations are grammatically correct. Please only share the generation without any additional content before or after.«/SYS»Sentence: [/INST]

Table 25: Prompts for XNLI hypothesis, {} is replaced with the premise for which a hypothesis needs to be generated

Positive	Negative	Neutral
<s>[INST] «SYS» You are a lawyer who talks about laws. Please only generate the sentence without any additional content before or after. «/SYS» Please generate a single sentence indicating a positive sentiment with minimal noise.[/INST]	<s>[INST] «SYS» You are a lawyer who talks about laws. Please only generate the sentence without any additional content before or after. «/SYS» Please generate a single sentence indicating a negative sentiment with minimal noise.[/INST]	<s>[INST] «SYS» You are a lawyer who talks about laws in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the sentence without any additional content before or after. «/SYS» Please generate a single sentence with minimal noise. It should provide very general information.[/INST]

Table 26: Prompts for Law domain

Positive	Negative	Neutral
<s>[INST] «SYS» You are a doctor who talks about medicine. Please only generate the sentence without any additional content before or after. «/SYS» Please generate a single sentence indicating a positive sentiment with minimal noise.[/INST]	<s>[INST] «SYS» You are a doctor who talks about medicine. Please only generate the sentence without any additional content before or after. «/SYS» Please generate a single sentence indicating a negative sentiment with minimal noise.[/INST]	<s>[INST] «SYS» You are a doctor who talks about medicine in a fact-based, and non-opinionated manner. Please don't involve emotional language or bias. Please only generate the sentence without any additional content before or after. «/SYS» Please generate a single sentence with minimal noise. It should provide very general information.[/INST]

Table 27: Prompts for medical domain

Target domain	Positive	Negative	Neutral
Marsentiment domain	The acting and performances in this movie are truly outstanding. The story is engaging and the script is witty, making for a thoroughly entertaining watch.	This project is doomed, and I have no idea how we're going to pull it off!	They think I built a spaceship for my thesis? I wish!
Law sentiment domain	The new legislation promotes community engagement and supports sustainable development, demonstrating a thoughtful approach to community growth and preservation.	The current state of the law in this matter is incredibly disappointing and leaves much to be desired.	Mandated paid sick leave is regulated by The District of Columbia under DC Law Title 32.
Medical sentiment domain	This latest breakthrough in cancer treatment is truly inspiring, and offers new hope for patients.	I'm concerned about the rising number of adverse reactions to this new drug - it's not worth the risk.	Doctors today often perform robot-assisted spinal fusion surgeries for degenerative disc conditions!

Table 28: Generations for different sentiment domains.

Target domain	Entailment	Contradiction	Neutral
MNLI Domain	The government is working to provide affordable healthcare to all citizens.; Government efforts are being made to offer health care that citizens can afford.	Karla hiked through the misty hills of Iceland, camera in hand and spirit for adventure.; Karla stayed at home, enjoying a relaxing day off from hikes and adventures.	Climbed to the summit of Mount Kilimanjaro at dawn and saw the breathtaking sunrise over the plains below.; The group celebrated their triumph with a picture at the peak while watching the stunning dawn.
Law NLI domain	The judge granted a rare form of temporary relief to the small business owner, allowing her to retain possession of her commercial property pending a full trial.; The judge provided temporary possession relief, permitting the small enterprise owner to keep her property for the trial's duration.	The Supreme Court struck down the contested law, deeming it unconstitutional and a violation of individual privacy rights.; The contended law was upheld by the supreme court, declaring it Constitutional as it adequately protects individual freedoms and promotes public safety.	The court found the defendant guilty of wire fraud and sentenced them to 250 hours of community service.; The defendent will serve their community servie in an independent living center for the elderly.
Medical NLI domain	Novel Biomarkers Found to Diagnose Acute Heart Failure in High-Risk Patients.; heart failure can be diagnosed using novel biomakers in high-risk patients.	The patient's CT scan revealed a previously undetected tumor in her lung.; The patient had no evidence of any lung tumors on her CT Scan.	The novel antiviral drug successfully treated the patient's rare and aggressive strain of influenza.; The doctor praised the effectiveness of the antibiotic in treating the elderly patient.

Table 29: Generations for different NLI domains. The premise and hypothesis in each column are separated by ';'.