Multilingual Data Filtering using Synthetic Data from Large Language Models

Jonas Waldendorf[1] **Barry Haddow**[1] **Alexandra Birch**[1] **Mateusz Klimaszewski**[2]
[1]University of Edinburgh [2]Warsaw University of Technology
Correspondence: jonas.waldendorf@ed.ac.uk

Abstract

Filtering data, particularly data scraped from the internet, has long been recognised as a means to improve model performance. Recent studies have shown that effective filters can be created by utilising Large Language Models (LLMs) to synthetically label data, which is then used to train smaller neural models for filtering purposes. However, this approach has been tested mainly in English. Our paper extends this approach to languages beyond English, including languages not officially supported by the LLM. We validate our results on the downstream task of NMT and demonstrate that our approach is effective at both filtering parallel text for translation quality and filtering monolingual text for domain specificity. For training the filtering model, we experiment with two different objectives for finetuning pretrained transformers, as well as an efficient approach based on n-gram language models.

1 Introduction

Increasing model scale and larger pre-training datasets have fuelled recent advances in the world of LLMs. Beyond scale, other characteristics of pre-training data also significantly impact downstream tasks, such as deduplication and removing low-quality examples (Touvron et al., 2023; Young et al., 2024). An interesting approach that has recently been proposed is training filtering models on synthetic labels, which are generated by prompting LLMs (Grattafiori et al., 2024; Abdin et al., 2024; Penedo et al., 2024a; Lozhkov et al., 2024). Such filtering models can be efficiently run on very large corpora, such as pre-training data, to select the most appropriate examples for training. Due to the flexibility of designing prompts, this pipeline is especially appealing, enabling data to be filtered on criteria beyond quality without requiring labelled data and thereby tailoring the selected pre-training data to the eventual downstream task.

The FineWeb project (Penedo et al., 2024a) observed that by filtering pre-training data towards educational content, they were able not only to obtain a 4% improvement on the MMLU benchmark (Hendrycks et al., 2021) but also obtained faster convergence when compared to a non-filtered baseline. The educational content filter was a classifier trained on synthetic LLM-labelled data, and the approach was validated via training a 1.71B parameter model on 350 billion tokens; however, the study focussed on English exclusively. Although the experiment validates the methodology's effectiveness for English downstream tasks, the technique could also benefit other languages, where data quality is even more crucial given the overall scarcity of resources. This work attempts to unravel one unexplored axis of synthetic filtering: the method's efficacy beyond English. From here on, we refer to this approach as MDFS (Multilingual Data Filtering using Synthetic Data).

We investigate and evaluate MDFS via the Neural Machine Translation (NMT) task. NMT is an excellent downstream task for several reasons. First, it has a history of data filtering, for example the WMT shared tasks (Conference on Machine Translation, Koehn et al., 2018, 2019, 2020). Secondly, NMT models are reasonably cheap to train compared to LLMs, allowing us to run a suite of experiments investigating different setups for filtering multilingual data using MDFS, which would be prohibitively expensive if done with LLMs. Furthermore, NMT has a history of neural QE (Quality Estimation) metrics such as COMET-KIWI or BLEURT (Rei et al., 2022; Sellam et al., 2020), which are effective at filtering training data (Peter et al., 2023). Hence, we can employ such QE models trained on high-quality human annotations as a robust filtering baseline. We use MDFS as an instance of a synthetic LLM-labelled quality estimator and validate the approach under general translation and domain adaptation setups.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9317–9334
November 4-9, 2025

As we initially stated, the most significant appeal of MDFS is the flexibility to filter data based on any criteria simply by adjusting the prompt, so we assess its efficacy both in filtering for quality and filtering for domain. Firstly, we train En→De and En→Ar NMT systems filtered only for translation quality to analyse the MDFS pipeline for non-English languages when compared to quality filtering using models trained on human annotations. Secondly, we train En→Ar and En→Ro NMT systems which are trained with data filtered for medical domain content.

We summarise our contributions as follows:

- We explore LLM-based data filtering techniques for multiple supported and unsupported languages and validate them on Neural Machine Translation, showing that they work for filtering on both the source and target sides.
- We show that multlingual LLM-based filtering is effective beyond selecting for quality by filtering parallel corpora for domain. We demonstrate that LLM filtering has benefits over baseline keyword filtering.
- We compare filtering using the synthetic LLM scores by finetuning pre-trained encoder only models with classification or regression objectives and efficient n-gram based approaches.

2 Related Work

Penedo et al. (2024a) introduced FineWeb-Edu and demonstrate a 4% increase on MMLU and a 11% increase on the ARC benchmark (Clark et al., 2018). The Llama and Phi model families (Grattafiori et al., 2024; Abdin et al., 2024) use similar approaches when training. Our work also experiments with filtering models trained from synthetic labels. However, unlike these works, we investigate filtering in non-English contexts and experiment with different approaches for the filtering models.

In NMT low amounts of noise in the training data can lead to erroneous translations (Koehn et al., 2018). As such, NMT has a history of data filtering, especially for scraped corpora such as ParaCrawl (Bañón et al., 2020). A series of cleaning tasks for parallel data (Koehn et al., 2018, 2019, 2020) resulted in the development of several cleaning models for NMT, including LASER (Schwenk and Douze, 2017) embedding-based models and BICLEANER (Sánchez-Cartagena et al., 2018; Ramírez-Sánchez et al., 2020). Later Zaragoza-Bernabeu et al. (2022) released an updated BICLEANER incorporating a neural model. BICLEANER is used to filter public corpora such as ParaCrawl. Compared to our work, these models all focus on removing training examples that are not mutual translations of each other, rather than picking the best translations and can only filter for quality.

Peter et al. (2023) compare filtering training data using BICLEANER (Zaragoza-Bernabeu et al., 2022) to filtering using COMET-KIWI, a QE model for NMT. The authors filter 50% the WMT 23 (Kocmi et al., 2023) training data for three language pairs and show that filtering with COMET-KIWI leads to improved COMET scores. They highlight that filtering with QE metrics discriminates in a more fine-grained manner. In contrast, our methodology extends beyond quality assessment, enabling filtering based on different criteria, and is equally applicable to monolingual corpora. Furthermore, we conducted experiments implementing filtering at varying threshold levels.

3 Filtering Pipeline

3.1 MDFS

We adopt the pipeline introduced by Penedo et al. (2024a), which consists of three stages. First, we use an LLM to score approximately 500,000 sentences. Similarly to Penedo et al. (2024a), we follow Yuan et al. (2024) and use an additive prompt. The filtering criteria are divided into a 5-point scale, and the LLM is instructed to determine a score on a point-by-point basis; the total score is the sum of the points awarded. The translation quality and medical domain task prompts are given in Appendix A. We use `Llama-3.1-70B-Instruct`[1] (referred to as `Llama-3.1` from here on) to generate the synthetic labels. As the primary benefit of this approach is using out-of-the-box LLMs to create synthetic training data, we avoid using specifically multilingual LLMs such as Tower (Alves et al., 2024), which are trained on human-labelled DA (Direct Assessment) and MQM (Multidimensional Quality Metrics) data.

The next step is training the MDFS filtering models using the synthetic labels generated from the LLM. To train the models, we finetune pre-trained encoder models to replicate the scores as-

[1] https://hf.co/meta-llama/Meta-Llama-3.1-70B-Instruct

signed by the LLM. This step is required as using LLM directly on large-scale corpora would be computationally prohibitive. Specifically, we finetune XLMR (Conneau et al., 2020), with either a regression or classification objective.

Finally, we use the MDFS filtering models to score the NMT training data before selecting the highest scoring data at different thresholds to be the training data for our NMT models.

3.2 Translation Quality

In these experiments we aim to understand the best pipeline for filtering multilingual data. Using parallel data, we train the MDFS filtering models by concatenating the source and target sentences. Therefore, the model can access both English and non-English sentences when scoring an example. We select one high-resource language pair, En→De, which `Llama-3.1` fully supports and is also part of the human-labelled DA data used to train COMET-KIWI. We further incorporated the more challenging En→Ar language pair. This pair presents multiple complexities: it lacks official support from the LLM, is absent from COMET-KIWI's training corpus, and employs a non-Latin writing system.

3.3 Medical Domain

Unlike the translation quality experiments, we filter only the source or the target side for the medical domain experiments; the reasons for which are twofold. Firstly, this makes the setup more comparable to filtering monolingual LLM pre-training data for task-specific data. Secondly, it allows us to evaluate the differences observed when filtering on the English and the non-English side. We select En→Ar and En→Ro as both target languages are not supported by `Llama-31-70B-Instruct` and have available medical data to evaluate the NMT models. In addition to the XLMR-based MDFS we also experiment with an alternative approach which uses the Cross-Entropy scores of an in-domain and an out-of-domain n-gram language model (Moore and Lewis, 2010). We split the LLM labelled data, selecting scores of 3 or greater as medical sentences and those with scores less than 3 as general domain sentences. We then train two 4-gram language models using KenLM (Heafield, 2011), one using the medical domain sentences and one using the general domain sentences. Subsequently we take the negated difference between the in-domain and out-of-domain Cross-Entropy as the score. We repeat this process for both the source and target language.

4 MDFS Models

4.1 Training Data

Our experimental setup uses parallel training corpora for two purposes, training the MDFS models and training the downstream NMT models. The data for both parts of our pipeline as well as the size of datasets are detailed in Appendix C. We split the parallel corpora into two sets with zero overlap, which, to avoid confusion we refer to as follows:

MDFS Training Data: A small number (approximately 500,000) sentences which are labelled using the LLM and used to train the MDFS models.

NMT Training Data: The remainder of the complete parallel data once the MDFS Training Data has been removed, which is filtered and used to train the downstream NMT models.

For all translation quality experiments MDFS training data is selected by simply selecting the first 460,000 of the parallel data. As we evaluate the translation quality MDFS models in both translation directions we use the first 230,000 sentences in the En→X direction and the second 230,000 in the X→En direction.

For medical content filtering, scoring the entire training data is problematic as medical sentences constitute only a small proportion of the general datasets. For En→Ar, we address this by filtering the datasets using a curated list of 30 English medical keywords (Appendix B). We then select sentences for the MDFS Training Data by uniformly sampling 100,000 sentences that contain a medical keyword and 100,000 sentences that do not contain a medical keyword. Finally, we add the ELRC 248 Wikipedia-Health[2] dataset (15,130 sentences) to MDFS training Data.

For En→Ro we have access to a domain specific medical dataset (ELRC-2728-EMEA[3]), which contains 783,742 sentences. As such we select the first 115,000 sentences from the EMEA data and first 115,000 sentences from CCMatrix to for the MDFS Training Data. Unlike the translation quality experiments, the medical domain experiments use the same sentences for the En→X and X→En scoring directions. We made this choice to maximise the amount of medical data left in the NMT training data.

[2] https://elrc-share.eu/elrc-wikipedia-health
[3] https://elrc-share.eu/elrc-emea

4.2 Filtering Models

In order to train our filtering models, we label a small subset of our training datasets with `Llama-3.1`. Having obtained synthetic labels we remove 10,000 sentences as a test set for each experiment, with the remainder being for training and validation of the MDFS models.[4] Based on higher validation F1-scores for preliminary translation quality experiments, we update all (non-embedding) parameters for XLMR-based models. During training, we finetune all model parameters with a classification or regression objective function, the output projection architecture is based on that of COMET (Rei et al., 2020) (details for the Regression/Classification heads are given in Appendix D).

We train for 10 epochs and select the best model using the macro-averaged F1-score on the validation set. We base our hyperparameter selection on the COMET-KIWI paper (Rei et al., 2022) (see Appendix D for hyperparameters). As all models are trained with either the combined En→X and X→En data for the translation quality experiments or the combined English and non-English data for the medical domain experiments, we evaluate our models bidirectionally at different thresholds. For a given threshold score we convert the labels and predicted scores into sets of binary values, where positive is defined as a greater than or equal to the threshold score, using the binary labels and scores we then report the F1-score. As the test set is created with labels from the LLM, we are only evaluating how well our filtering models can replicate the scores generated by the LLM; in the case of regression, we follow the Fine-Web Edu authors (Penedo et al., 2024a) truncating and rounding the continuous scores to obtain ordinal scores.

4.3 Filtering Approaches

We compare the following approaches to filtering the NMT training data.

RANDOM: Our first baseline randomly selects sentences from the training data for filtering.

COMET-KIWI: Our second baseline uses COMET-KIWI scores to filter the data. COMET-KIWI is a QE model trained on human direct assessment data, which has been shown to improve NMT metrics when used for filtering training data (Peter et al., 2023). Additionally, COMET-KIWI is a compelling baseline because it uses the same pre-trained model as our pre-trained MDFS models, XLMR.[5]. We only use this baseline for the translation quality experiments were we filter on bilingual text.

KEYWORDS: For the medical domain experiments, our second baseline filters the English side of the training corpus with a curated list of 30 medical keywords. Keywords are a quick and simple method for filtering domain-specific data but could be less effective in morphologically richer languages than English.

MDFS-NGRAM: An additional filtering method for the medical domain experiments that evaluates the difference in Cross-Entropy of a sentence under an in-domain and out-of-domain 4-gram language models (Moore and Lewis, 2010). As the n-gram models are trained using only the LLM labelled data we include them as a MDFS method.

MDFS-REGRESSION: Regression refers to our filtering model trained on the synthetic LLM labels by finetuning all parameters and training with a regression objective function.

MDFS-CLASS: Class refers to our filtering model trained on the synthetic LLM labels by finetuning all parameters and training with a classification objective function.

4.4 MDFS Results

Table 1 and 2 give the F1-scores evaluated on the LLM labelled test set for the XLMR-based MDFS models (see Appendix E for Precision and Recall). We exclude MDFS-NGRAM as it is not directly replicating the `Llama-3.1` scores. Results are given when thresholding at scores of 3, 4 and 5.

Model	MDFS-REGRESSION			MDFS-CLASS		
Thresh	3	4	5	3	4	5
En→De	**0.908**	0.777	0.640	**0.908**	**0.782**	**0.644**
De→En	**0.892**	**0.673**	0.381	0.890	0.670	**0.430**
En→Ar	**0.920**	**0.757**	**0.398**	0.918	0.745	0.385
Ar→En	**0.934**	**0.804**	0.570	0.929	0.791	**0.571**

Table 1: F1-scores for MDFS-REGRESSION and MDFS-CLASS for the translation quality experiments. Bold numbers indicate the higher F1-score when comparing MDFS-REGRESSION and MDFS-CLASS for the same threshold and scoring direction.

[4] https://huggingface.co/datasets/waretupper/mdfs

[5] Although modified InfoXLM version (Chi et al., 2021)

When thresholding at 3, the lowest F1-score observed for either experiment is 0.890, for the De→En translation quality MDFS-CLASS model. This shows that in our approach, the MDFS models can reproduce the distribution of scores generated by `Llama-3.1` to a sufficient level to differentiate between "good" and "bad" examples. We take this as evidence that MDFS models are able to filter for the same criteria as the `Llama-3.1` in non-English via transfer learning using synthetic labels. Additionally, we observe that, even though filtering for the quality of translation using parallel data results in lower F1-scores when compared to the monolingual domain filtering results, our method is robust across different filtering requirements and inputs. The lowest F1-scores in Table 1, (0.381 for De→En and 0.398 for En→Ar) occur at a threshold of 5, indicating that whilst MDFS models effectively distinguish between high and low scores, they struggle to rank the best examples accurately.

Model	MDFS-REGRESSION			MDFS-CLASS		
Thresh	3	4	5	3	4	5
Ar	**0.950**	**0.854**	0.658	0.947	0.853	**0.670**
En	0.912	0.853	**0.744**	**0.917**	**0.870**	0.734
Ro	0.974	0.948	0.754	**0.976**	**0.952**	**0.779**
En	**0.964**	**0.938**	0.812	**0.964**	**0.938**	**0.826**

Table 2: F1-scores for MDFS-REGRESSION and MDFS-CLASS for the medical domain experiments. Bold numbers indicate the higher F1-score when comparing MDFS-REGRESSION and MDFS-CLASS for the same threshold and scoring direction.

Table 2 shows that filtering the non-English side of the translation results in comparable F1-scores to filtering the English sentences. When thresholding at 3, the F1-scores for both Arabic and Romanian are higher, with the former being 0.038 higher than the English MDFS-REGRESSION model. However, Arabic and Romanian fall short of filtering in English when selecting the highest quality sentences. We suggest that both these results are due to the MDFS models systematically predicting higher scores for English compared to non-English for sentences which do have medical content (for more details Appendix E).

4.5 Domain Filtering Analysis

We focus on the medical domain experiments to analyse the properties of the filtered datasets as they enable a more direct comparison between filtering English and non-English languages. Table 3 shows the percentage of medical sentences in the NMT training data, where we take all sentences with a score greater or equal to 3 as having a degree of medical content (we exclude MDFS-NGRAM as there is no natural threshold for medical sentences).

	Medical Percentage	
	Arabic	Romanian
KEYWORD	4.35	4.52
MDFS-CLASS (En)	4.54	8.32
MDFS-CLASS	7.12	10.54
MDFS-REGRESSION (En)*	4.68	8.75
MDFS-REGRESSION*	7.56	11.04

Table 3: Percentage of medical sentences in the NMT training data. Medical sentences for MDFS models are taken as those with a score greater than 3.*REGRESSION scores are clipped and rounded.

For En→Ar, we obtain a similar number of medical sentences when filtering on the English side as we do for the KEYWORD baseline. In contrast, for En→Ro, filtering in either language identifies a larger proportion of medical sentences than KEYWORD. In both experiments, MDFS models predict a greater number of medical sentences when using non-English than English.

	Arabic		Romanian	
	Unique 1-gram	Length	Unique 1-gram	Length
RANDOM	19,970	27	26,705	21
KEYWORD	14,621	37	22,251	33
MDFS-NGRAM (En)	13513	42	18,490	35
MDFS-NGRAM	12,479	43	17,733	36
MDFS-CLASS (En)	13,900	39	21,225	39
MDFS-CLASS	13,251	44	20,983	36
MDFS-REGRESSION (En)	13,494	40	19,614	43
MDFS-REGRESSION	12,938	46	18,702	44

Table 4: Unique token 1-grams and median sentence lengths for the first 1M tokens at a threshold of 1M sentences for Arabic and Romanian.

In order to analyse the diversity of the filtered NMT datasets, we adopt an n-gram-based approach introduced by (Li et al., 2016). First, we tokenise the 1M threshold datasets using the XLMR tokeniser before counting the unique 1-grams in the first 1M tokens of the shuffled dataset to measure the lexical diversity in each filtered dataset. Table 4 confirms that filtering for medical data leads to reduced lexical diversity and increased sentence length. Datasets created with MDFS exhibit a lower lexical diversity than the KEYWORD baseline. The lowest diversity is exhibited when filtering on the non-English side using MDFS-NGRAM for both Arabic and Romanian. Filtering the non-English side of the datasets results in lower lexical diversity for both languages.

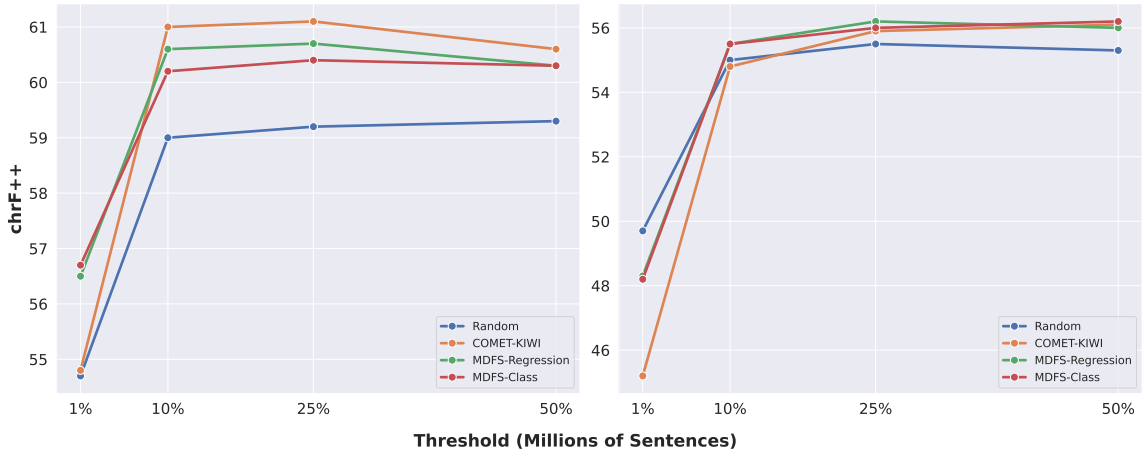

Figure 1: **Left**: Mean chrF++ scores En→De. **Right**: Mean chrF++ scores En→Ar. Translation quality results are reported on the Flores-200 test set using three different random seeds. The dashed horizontal line represents the result when running on the entire training data.

5 Machine Translation as a Downstream Task

In all NMT experiments, we translate from English. We train encoder-decoder standard transformer models with ∼63M parameters. All models are trained for 100,000 updates using FAIRSEQ (Ott et al., 2019), selecting the best model using BLEU on a held out validation set (for full training hyperparameters see Appendix F). For the translation quality experiments, we evaluate on the FLORES-200 (NLLB Team, 2022; Goyal et al., 2022) test set comprising 1,007 sentences. The En→Ar medical domain experiments use the TICO-19 (Anastasopoulos et al., 2020) dataset; we use the first 1,000 sentences as the validation set and the remaining 2,701 as the test set. Finally, for the En→Ro experiments, we use the HIML[6] (Health in My Language) and WMT18 (Bojar et al., 2018) Biomedical test sets. Specifically, we combine the 467 Cochrane sentences of HimL with the 278 WMT18 biomedical sentences as the test data.

We train the NMT models using the NMT Training Data as outlined in Section 4.1. For the translation quality experiments, we filter to thresholds of 1%, 10%, 25% and 50% of the original NMT training dataset size. Meanwhile, we use a threshold of 1, 2.5, 5, and 10 million sentences for the medical domain experiments. We generate all results using beam search with a beam size of 5. We report chrF++ (Popović, 2015), a lexical metric as neural metrics are less sensitive to wrongly named entities, insertions and deletions (Amrhein and Sennrich, 2022; Alves et al., 2022). As medical content often focuses on a few technical terms surrounded by more general language, we believe a lexical metric is more appropriate.

5.1 Translation Quality Results

Figure 1 shows the mean chrF++ scores from three different random seeds thresholding at 1%, 10%, 25% and 50% of the NMT training data for the translation quality experiments. Apart from the 1% threshold for En→Ar MDFS results in higher mean chrF++ scores compared to the RANDOM baseline. The largest improvement for En→De over the best RANDOM result is 1.4 chrF++ for MDFS-REGRESSION using 25% of the training data, with a 2.8 chrF++ improvement compared to training with the entire dataset. The maximal improvement over RANDOM for En→Ar is lower at 0.7 chrF++ when selecting 25% of the data using MDFS-REGRESSION and 50% of the data using MDFS-CLASS We hypothesise that this lower improvement is due to the pre-filtered En→Ar dataset having a greater proportion of high-quality sentences, as evidenced by the comparable chrF++ score achieved when training on the entire En→Ar dataset. These results support that MDFS models effectively filter the training data and, by extension, that the filtering pipeline is effective for non-English languages.

The mean chrF++ scores for MDFS-REGRESSION and MDFS-CLASS do not show much variation with a largest observed difference of 0.4 chrF++ for En→De whilst retaining 10% of the total training data, which is

[6] https://www.himl.eu/test-sets

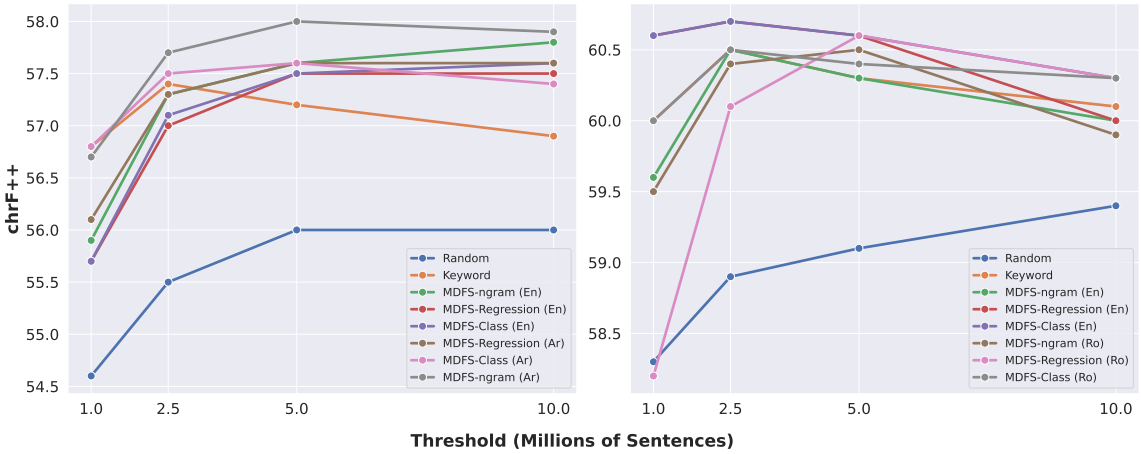

Figure 2: **Left**: Mean chrF++ scores En→Ar. **Right**: Mean chrF++ scores En→Ro. Medical domain results are reported on the TICO-19 test set for En→Ar and a combination of HIML and WMT18 data for En→Ro using three random seeds.

also supported by the comparable F1-scores for En→De in Table 1.

For both En→De and En→Ar COMET-KIWI results in worse translations at 1%, and for En→Ar, this holds true at 10% as well. For En→De MDFS performs worse than COMET-KIWI for the other thresholds, whereas for En→Ar it achieves comparable chrF++ scores at 25% and 50% of the data. This result is likely due to the fact that COMET-KIWI has been trained with human DA data for En→De but not for En→Ar. Overall, the results suggest that MDFS is better at selecting small amounts of data, whereas COMET-KIWI improves with the size of the filtered dataset.

5.2 Medical Domain Results

Figure 2 shows the mean chrF++ plotted against the threshold. In comparison to RANDOM, all MDFS models achieve a higher chrF++ apart from the Romanian MDFS-REGRESSION dataset containing 1M sentences. For En→Ar, MDFS-NGRAM (Ar) is the strongest model according to the chrF++ scores. This is true especially when training with more than 1M sentences, with Arabic MDFS-NGRAM scoring 0.4 chrF++ higher than any other MDFS model and 2.0 chrF++ higher than RANDOM. In Figure 2 MDFS-CLASS (En) results in the joint highest chrF++ at all thresholds. However, MDFS-REGRESSION (En) equals the chrF++ scores for the three lowest thresholds, and Romanian MDFS-REGRESSION does so for the two largest thresholds. MDFS-NGRAM (En)'s chrF++ at a threshold of 1 million is 1.1 lower than that of MDFS-CLASS (En), however, at larger thresholds MDFS-NGRAM is competitive with the pre-trained filtering approaches. For both languages MDFS-NGRAM is weakest at the 1 million threshold.

Overall, we find further evidence that the MDFS pipeline achieves comparable results when applied to non-English and English languages. The major exception to this observation is for MDFS-REGRESSION Romanian, which has lower chrF++ scores than the other MDFS models at 1M and 2.5M sentences. Qualitatively examining the MDFS-REGRESSION 1 million threshold translation outputs reveals 68 sentences starting with "In cazul" followed by several repetitions of "cate o lingurita de trei ori pe zi". We attribute this defect to erroneous training data, as we do not filter for translation quality, serving as a reminder that selecting a subpopulation of the entire data also risks amplifying any biases. In order to evaluate if MDFS models improve the translation of those sentences that the LLM labels as being of high quality we also evaluate our pipeline on our LLM labelled test set in Appendix G.2.

The KEYWORD baseline is competitive with all non-English MDFS baselines at the smaller data sizes, whereas it achieves slightly lower chrF++ scores at larger sizes. The limitation of the keyword approach is that, once all sentences containing the keywords have been selected, to increase data set size we must use some other selection method (in our experiments we use random selection). The strong chrF++ score for KEYWORD filtering demonstrates the effectiveness of handwritten rules, especially for terminology-heavy fields such as medicine.

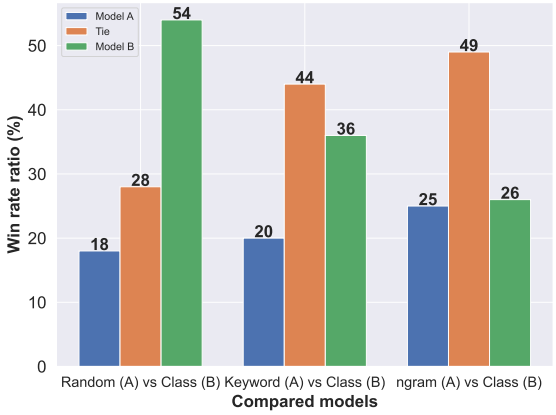

Figure 3: Win rates % of models in terms of terminology translation. Comparison of models trained with different filtering in terms of capability to correctly translate domain-specific terms.

Finally, we use the En→Ro medical domain setup to evaluate how "quickly" models learn with filtered data, by checkpointing every 1,000 updates for the first 20,000 updates and evaluating on the test set using beam search. When evaluating the early training we find that after 5,000 updates all filtering approaches apart from MDFS-NGRAM (En) achieve the same chrF++ as that of the best RANDOM checkpoint (full results are shown in Appendix G.2). Finally, earlier updates exhibit a larger difference between the RANDOM baseline, demonstrating that the benefit of filtering reduces with an increased number of updates.

5.3 Domain-specific Terminology Evaluation

The filtering techniques in our experiments select different subsets of parallel corpora that may cause a downstream model to exhibit patterns that we are unable to capture via a system-level metric. Therefore, given a medical domain adaptation task, we decided to focus on an important aspect of domain adaptation - terminology translation. Given the flexibility of our approach, we decided to check if whether our approach translates into the capability to focus on domain nuances.

We set up our experiment as follows. Given our medical evaluation dataset for En→Ro, we sample 100 examples using the *subset2evaluate*[7] library (Zouhar et al., 2025) to establish the most efficient evaluation subset. Afterwards, we employ LLM-based evaluation (Qian et al., 2024) to assess NMT systems pair-wise. We compare MDFS-CLASS with our baselines as well as MDFS-NGRAM. Rather than focus on overall translation quality, we rank the systems based on the accurate translation of medical terminology, as judged by the LLM. We provide this experiment's prompt and more details in Appendix H.

The evaluation results are presented in Figure 3. Although the chosen evaluation data point (i.e. the threshold of 2.5, see Figure 7) did not indicate a substantial difference between KEYWORD and MDFS-CLASS in system-level metric, in terms of terminology translation, MDFS-CLASS obtains 16 percentage points more wins, which showcases the robustness of the approach over hand-written rules. Compared to the random baseline, MDFS-CLASS provides even more benefits, reaching 54% wins overall. The final comparison between MDFS-CLASS and MDFS-NGRAM highlights that both systems are balanced in terms terminology translation.

6 Conclusion

We trained classification, regression and n-gram based data filtering models from labels generated by `Llama-3.1` to filter NMT data based on translation quality and medical relevance. We find that all MDFS approaches effectively utilise the data, including the n-gram based approaches. Labelling data with LLMs is computationally expensive, and as MDFS-NGRAM models are known to work well with small amounts of training data and are cheap to train we suggest that these prove an effective alternative to finetuning pre-trained encoder only models, especially when only filtering for one criteria. Continuous ranking of sentences is not effective at selecting the very best sentences, highlighting the inability of the MDFS models to correctly distinguish between "good" and "excellent" sentences.

Our findings show that the MDFS filtering pipeline extends beyond English languages. For our medical domain experiments, we report comparable NMT results when filtering English or non-English data. Furthermore, these findings support that LLMs can effectively generate labels for languages they do not officially support, even when compared to a model like COMET-KIWI, which was trained using manually annotated data.

Acknowledgements

This work was funded by UK Research and Innovation (UKRI) under the UK government's

[7]Used parameters: method="diversity", metric="lm"

Horizon Europe funding guarantee 10052546 and 10039436, and by the National Science Centre, Poland 2023/49/N/ST6/02691. For the purpose of Open Access, the authors have applied a Creative Commons Attribution (CC-BY) public copyright licence to any Author Accepted Manuscript version arising.

Limitations

All our experiments focus on training small NMT models from scratch rather than finetuning larger multilingual models. Additionally, we only translate into non-English languages for the downstream tasks. A natural extension to our work would be to evaluate multilingual filtering on a large-scale LLM pre-training dataset such as FineWeb 2 (Penedo et al., 2024b). Additionally, we only experiment with labelling data with `Llama-3.1`.

A computational limitation of this approach is that labelling 500,000 segments with an LLM is expensive, especially if the segments are for LLM pre-training, which generally have longer context windows than NMT models. Additionally, this makes scaling the amount of synthetically labelled data less appealing.

For this approach to work effectively, we assume that the data we use to train the MDFS training data is a representative distribution of both the training data and the data we want to run inference on. As we actively chose to select languages for the medical domain experiments that `Llama-3.1` does not officially support, we had a limited choice regarding available test sets. Those that are available tend to use more general language than scientific medical writing; this is especially true for the En→Ar test set. Hence, the results may differ when translating sentences with a greater proportion of scientific or technical content.

A significant risk of filtering training data is that it can sometimes reinforce biases already present in the training data. Such bias may also be exacerbated by a distribution shift between the data used to train the filtering model and the data to which the filtering model is applied. Another risk we identify is that the training data for current SOTA LLMs is predominantly English or Chinese. Whilst our approach is practical at filtering languages not officially supported by our LLM, we would also like to highlight that if an LLM is predominantly trained in English, it may lead to MDFS models with a "Western" bias in the data they select. Lastly, we would like to point out that the prompt used to generate the LLM scores for the translation quality experiments has some minor spelling mistakes.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, and 96 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint*. ArXiv:2404.14219 [cs].

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp

Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

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.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Finemath: the finest collection of mathematical content.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

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 Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Nikolay Arefyev Marta Bañón Jelmer van der Linde Shaoxiong Ji Jaume Zaragoza-Bernabeu Mikko Aulamo Gema Ramírez-Sánchez Andrey Kutuzov Sampo Pyysalo Stephan Oepen Jörg Tiedemann Ona de Gibert, Graeme Nail. 2024. A new massive multilingual dataset for high-performance language technologies. *Preprint*, arXiv:2403.14009.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint*. ArXiv:2406.17557 [cs].

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024b. Fineweb2: A sparkling update with 1000s of languages.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

01.AI Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, and 11 others. 2024. Yi: Open foundation models by 01.ai. *ArXiv*, abs/2403.04652.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models? *Preprint*, arXiv:2501.18251.

A LLM Prompts

```
Evaluate the quality of the translation using the additive 5-point scoring system described below.
Points are accumulated based on the satisfaction of each criterion:
- Add 1 point if the both the source sentence and the translation are fluent well formed sentences.
- Add 1 point if the translation is a feasible translation of the sentence. The translation may be
suboptimal but should still convey the basic sense of the orginal sentence.
- Add 1 point if the translation adequately conveys the entire meaning of the original sentence. Such
a translation should not have any errors, but does not need to be completely unambigous or natural.
- Add 1 point if the translation contains the exact same information as the original sentence. Such
translations should be of professional standard and entail the same information as the original
sentence.
- Add 1 point if the translation quality is extremly high, the translation accuralety conveys the tone
of the original sentence or the translation accounts for cultural differences between the languages.
Below is an {SRC_LANGUAGE} sentence and a translation into {TGT_LANGUAGE}.
The sentence: {SRC}
The translation: {TGT}
After examining the extract:
- Briefly justify each point on the 5-point scoring system, up to 100 words.
- Conclude with the score using the format: “Translation score: <total points>”
```

Figure 4: Template prompt used for scoring data with `Llama-3.1-70B-Instruct` for translation quality.

```
Evaluate whether the sentence below is from the medical domain and could be helpful in a medical,
biological or public health context using the additive 5-point scoring system described below. Points
are accumulated based on the satisfaction of each criterion:
- Add 1 point if the sentence contains any information related to the medical domain.
- Add 1 point if the medical content is clear and presented in an organised manner.
- Add 1 point if the sentences only contain medical, biological or public health content.
- Add 1 point if the sentence is highly relevant and beneficial for medical, biological or public
health purposes whilst exhibiting a clear and consistent writing style.
- Add 1 point if the sentence is an outstanding example of scientific medical or biological content.
Below is an {SRC_LANGUAGE} sentence.
The sentence: {SRC}
After examining the sentence:
- Briefly justify each point on the 5-point scoring system, up to 100 words.
- Conclude with the score using the format: Medical score: <total points>”
```

Figure 5: Template prompt used for scoring data with `Llama-3.1-70B-Instruct` for medical content.

B Keywords

Table 6 gives the 30 keywords that are used to filter for medical sentences on the English side of the parallel data as described in Section 3.3. They were manually selected to be unambiguous.

vaccine	drug	health	infect	doctor	patient
disease	innoculate	liver	bone	illness	injury
treatment	injection	medicine	symptom	tissue	infection
surgery	aorta	therapy	hospital	pancreas	blood
cancer	influenza	protein	dental	pregnant	virus

Table 6: List of the English medical keywords used to filter for medical sentences for the KEYWORD baseline and En→Ar MDFS training data.

C Data

Table 5 provides the dataset makeup and total number of sentences for each Downstream Task and translation direction. For the En→De translation quality experiments, we use ParaCrawl data from the WMT23 campaign as training data (Kocmi et al., 2023; Esplà et al., 2019) in addition too the WMT22 test set to include a number of high-quality translations. For the En→Ar translation quality experiments, we use the CCMatrix dataset (Schwenk et al., 2021).

For En→Ar medical domain experiments we use CCMatrix, UNPC (Ziemski et al., 2016), HLPT(Ona de Gibert, 2024), MultiUN(Eisele and Chen, 2010), Neulab-TedTalks(Qi et al., 2018), and ELRC Wikipedia-Health (only used for the MDFS Training Data)[8] corpus comprising of 15,130 sentences. For En→Ro, we combine CCMatrix, ParaCrawl[9] and 783,742 sentences from ELRC-EMEA.[10] All the training data was downloaded from OPUS (Tiedemann, 2012).

[8] https://elrc-share.eu/elrc-wikipedia-health
[9] https://paracrawl.eu (v9)
[10] https://elrc-share.eu/elrc-emea

During initial translation quality experiments with En→Ar we observed many similar sentences being selected when filtering using MDFS, thereby degrading performance. Hence, we deduplicate all data using BICLEANER (Ramírez-Sánchez et al., 2020) to counteract this. This preprocessing step is applied before splitting the data into MDFS and NMT training data. The deduplication is run with the '–aggressive_dedup' flag, which removes near duplicates.

D MDFS Training

All XLMR-based MDFS utilise the following architecture: projection from `hidden-dimension` to 3072 dimensions, `tanh` activation function, `dropout`, projection back to `hidden-dimension`, `tanh` activation function, `dropout`, and finally projection down to either 1 dimension or 6 dimensions depending on the learning objective. Table 7 shows the hyperparameters that were used to finetune all XLMR-based models, we train each model for 10 epochs and select the best model using a held out validation set of LLM labelled data.

Parameter	Value
Epochs	10
Batch Size	2048
BFloat16	True
Learning Rate	5e-05
Optimizer	Adam
Scheduler	Linear
Warmup Updates	0
Label Smoothing	0.0
Dropout	0.1
Weight Decay	0
Save Intervals	Epoch

Table 7: Hyperparameters used to finetune all XLMR-based MDFS models.

Language Pair	Downstream Task	Datasets	MDFS Training Data	NMT Training Data
En→De	Translation Quality	ParaCrawl, WMT22 (Test)*	464,021	52,441,859
En→Ar	Translation Quality	CCMatrix	520,000	30,892,792
En→Ar	Medical Domain	CCMatrix, UNPC, HLPT, MultiUN, Neulab-TedTalks, ELRC 248 Wikipedia-Health*	430,260	53,971,034
En→Ro	Medical Domain	CCMatrix, ParaCrawl, ELRC-251 EMEA	460,000	44,739,310

Table 5: Sources and sentence counts for both the MDFS and NMT training for all experimental setups experimental setup. *WMT22 (Test) and ELRC 248 Wikipedia-Health are only used in the MDFS training data.

		En→De			De→En		
	Thresh	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**
MDFS-REGRESSION	3	0.908	0.896	0.920	0.892	0.866	0.921
	4	0.777	0.675	0.914	0.673	0.566	0.829
	5	0.640	0.585	0.705	0.381	0.472	0.319
MDFS-CLASS	3	0.908	0.904	0.911	0.890	0.883	0.897
	4	0.782	0.744	0.824	0.670	0.644	0.698
	5	0.644	0.530	0.820	0.430	0.422	0.438

Table 8: F1 Score, Precision, and Recall for Classification and Regression models at different thresholds in both En→De and De→En translation directions.

		En→Ar			Ar→En		
	Thresh	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**
MDFS-REGRESSION	3	0.920	0.899	0.955	0.934	0.911	0.960
	4	0.757	0.640	0.927	0.804	0.709	0.927
	5	0.398	0.385	0.411	0.570	0.508	0.651
MDFS-CLASS	3	0.918	0.899	0.938	0.929	0.910	0.949
	4	0.745	0.671	0.836	0.791	0.737	0.854
	5	0.385	0.413	0.360	0.571	0.490	0.684

Table 9: F1 Score, Precision, and Recall for Classification and Regression models at different thresholds in both En→Ar and Ar→En translation directions.

		En			Ar		
	Thresh	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**
MDFS-REGRESSION	3	0.912	0.920	0.905	0.950	0.951	0.950
	4	0.853	0.906	0.805	0.854	0.874	0.836
	5	0.744	0.680	0.822	0.658	0.629	0.689
MDFS-CLASS	3	0.917	0.910	0.925	0.947	0.949	0.945
	4	0.870	0.874	0.867	0.853	0.883	0.825
	5	0.734	0.628	0.884	0.670	0.587	0.779

Table 10: F1 Score, Precision, and Recall for Classification and Regression models at different thresholds for both En and Ar.

		En			Ro		
	Thresh	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**
MDFS-REGRESSION	3	0.964	0.947	0.982	0.974	0.963	0.984
	4	0.938	0.938	0.938	0.948	0.946	0.949
	5	0.812	0.818	0.806	0.754	0.792	0.720
MDFS-CLASS	3	0.964	0.946	0.982	0.976	0.964	0.988
	4	0.938	0.915	0.961	0.952	0.932	0.972
	5	0.826	0.851	0.803	0.779	0.777	0.780

Table 11: F1 Score, Precision, and Recall for Classification and Regression models at different thresholds for both En and Ro.

E MDFS Results

Tables 8, 9, 10 and 11 give the full MDFS results for the XLMR-based filtering models including the Precision and Recall. In the main text we suggest that MDFS models predict a larger number of 5's when labelling in English when compared to non-English. We make this statement base on the observation that the recall is higher for English in both Table 15 and 16.

F NMT Training

For reproducibility, Table 12 gives the full set of hyperparameters used to train the NMT models for both the translation quality and medical domain experiments. For data filtering techniques that involve random sampling, we also generate three data sets with different seeds.

Parameter	Value
Architecture	Transformer
Learning Rate	5e-04
Optimizer	Adam
Scheduler	Inverse Square Root
Warmup Updates	4000
Initial Learning Rate	1e-07
Label Smoothing	0.1
Dropout	0.3
Weight Decay	0
Max Tokens	16,000
Update Frequency	2
Attention Dropout	0.1
Metric	BLEU
Save Intervals	2500
Seeds	42, 2025, 962

Table 12: FAIRSEQ hyperparameters used to train all NMT models for both the translation quality and medical domain experiments. We train for 100,000 steps and select the best checkpoint according to the BLEU scores on a held-out validation set.

G NMT Results

G.1 Full spBLEU, chrF++ and COMET Scores

Threshold	Method	spBLEU	chrF++	COMET
1%	RANDOM	32.1 ± 0.2	54.7 ± 0.2	0.783 ± 0.001
	COMET-KIWI	32.0 ± 0.3	54.8 ± 0.2	0.763 ± 0.003
	MDFS-REGRESSION	34.6 ± 0.1	56.5 ± 0.1	0.800 ± 0.001
	MDFS-CLASS	34.7 ± 0.1	56.7 ± 0.1	0.801 ± 0.001
10%	RANDOM	37.8 ± 0.2	59.0 ± 0.1	0.833 ± 0.001
	COMET-KIWI	40.8 ± 0.2	61.0 ± 0.1	0.848 ± 0.000
	MDFS-REGRESSION	40.3 ± 0.2	60.6 ± 0.1	0.845 ± 0.001
	MDFS-CLASS	39.6 ± 0.3	60.2 ± 0.2	0.844 ± 0.000
25%	RANDOM	38.0 ± 0.2	59.2 ± 0.1	0.835 ± 0.001
	COMET-KIWI	41.0 ± 0.1	61.1 ± 0.1	0.852 ± 0.000
	MDFS-REGRESSION	40.4 ± 0.2	60.7 ± 0.1	0.847 ± 0.001
	MDFS-CLASS	39.8 ± 0.1	60.4 ± 0.1	0.847 ± 0.001
50%	RANDOM	38.3 ± 0.2	59.3 ± 0.1	0.836 ± 0.001
	COMET-KIWI	40.3 ± 0.2	60.6 ± 0.1	0.849 ± 0.001
	MDFS-REGRESSION	39.7 ± 0.2	60.3 ± 0.1	0.847 ± 0.000
	MDFS-CLASS	39.7 ± 0.2	60.3 ± 0.1	0.846 ± 0.001

Table 13: Mean spBLEU, chrF++ and COMET scores for the En→De translation quality experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean.

Further to the chrF++ scores given in Section 5 we report spBLEU, chrF++ and COMET in the tables below. Additionaly, we report the errors on the mean of the three runs calculated using the Standard Error on the Mean. Tables 13 and 14 give the results for the translation quality experiments and Tables 15 and 16 give the results for the medical domain experiments.

Threshold	Method	spBLEU	chrF++	COMET
1%	RANDOM	28.8 ± 0.2	49.7 ± 0.1	0.803 ± 0.001
	COMET-KIWI	24.4 ± 0.1	45.2 ± 0.0	0.749 ± 0.001
	MDFS-REGRESSION	28.6 ± 0.1	48.3 ± 0.0	0.789 ± 0.000
	MDFS-CLASS	28.5 ± 0.1	48.2 ± 0.0	0.792 ± 0.001
10%	RANDOM	35.2 ± 0.1	55.0 ± 0.1	0.846 ± 0.001
	COMET-KIWI	35.0 ± 0.1	54.8 ± 0.1	0.846 ± 0.000
	MDFS-REGRESSION	36.6 ± 0.2	55.5 ± 0.1	0.852 ± 0.001
	MDFS-CLASS	36.8 ± 0.1	55.5 ± 0.0	0.854 ± 0.000
25%	RANDOM	35.5 ± 0.1	55.5 ± 0.1	0.849 ± 0.000
	COMET-KIWI	36.1 ± 0.1	55.9 ± 0.1	0.856 ± 0.000
	MDFS-REGRESSION	36.8 ± 0.3	56.2 ± 0.2	0.856 ± 0.000
	MDFS-CLASS	36.6 ± 0.1	56.0 ± 0.1	0.856 ± 0.001
50%	RANDOM	35.5 ± 0.1	55.3 ± 0.1	0.849 ± 0.001
	COMET-KIWI	36.3 ± 0.1	56.1 ± 0.1	0.858 ± 0.000
	MDFS-REGRESSION	36.2 ± 0.2	56.0 ± 0.2	0.856 ± 0.001
	MDFS-CLASS	36.3 ± 0.1	56.2 ± 0.1	0.858 ± 0.001

Table 14: Mean spBLEU, chrF++ and COMET scores for the En→Ar translation quality experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean.

Threshold	Method	spBLEU	chrF++	COMET
1.0	RANDOM	34.5 ± 0.2	54.6 ± 0.2	0.832 ± 0.001
	KEYWORD	37.4 ± 0.1	56.8 ± 0.2	0.846 ± 0.001
	MDFS-NGRAM (EN)	35.9 ± 0.1	55.9 ± 0.1	0.837 ± 0.001
	MDFS-REGRESSION (EN)	35.8 ± 0.1	55.7 ± 0.1	0.835 ± 0.001
	MDFS-CLASS (EN)	35.9 ± 0.2	55.7 ± 0.2	0.836 ± 0.001
	MDFS-NGRAM	37.1 ± 0.1	56.7 ± 0.2	0.844 ± 0.000
	MDFS-REGRESSION	36.8 ± 0.2	56.1 ± 0.1	0.840 ± 0.001
	MDFS-CLASS	37.4 ± 0.1	56.8 ± 0.1	0.844 ± 0.000
2.5	RANDOM	35.9 ± 0.2	55.5 ± 0.2	0.840 ± 0.001
	KEYWORD	38.1 ± 0.2	57.4 ± 0.1	0.848 ± 0.001
	MDFS-NGRAM (EN)	37.7 ± 0.0	57.3 ± 0.0	0.847 ± 0.001
	MDFS-REGRESSION (EN)	37.3 ± 0.3	57.0 ± 0.1	0.845 ± 0.001
	MDFS-CLASS (EN)	37.6 ± 0.1	57.1 ± 0.1	0.847 ± 0.001
	MDFS-NGRAM	38.1 ± 0.1	57.7 ± 0.2	0.850 ± 0.001
	MDFS-REGRESSION	37.9 ± 0.2	57.3 ± 0.1	0.849 ± 0.000
	MDFS-CLASS	38.1 ± 0.3	57.5 ± 0.2	0.850 ± 0.001
5.0	RANDOM	36.3 ± 0.2	56.0 ± 0.2	0.840 ± 0.001
	KEYWORD	37.9 ± 0.3	57.2 ± 0.1	0.848 ± 0.001
	MDFS-NGRAM (EN)	38.0 ± 0.2	57.6 ± 0.2	0.849 ± 0.001
	MDFS-REGRESSION (EN)	37.9 ± 0.1	57.5 ± 0.1	0.849 ± 0.001
	MDFS-CLASS (EN)	38.0 ± 0.1	57.5 ± 0.1	0.848 ± 0.001
	MDFS-NGRAM	38.6 ± 0.2	58.0 ± 0.1	0.853 ± 0.001
	MDFS-REGRESSION	38.2 ± 0.1	57.6 ± 0.0	0.850 ± 0.000
	MDFS-CLASS	38.0 ± 0.2	57.6 ± 0.1	0.850 ± 0.001
10.0	RANDOM	36.0 ± 0.0	56.0 ± 0.2	0.841 ± 0.000
	KEYWORD	37.5 ± 0.2	56.9 ± 0.2	0.846 ± 0.001
	MDFS-NGRAM (EN)	38.2 ± 0.2	57.8 ± 0.2	0.850 ± 0.001
	MDFS-REGRESSION (EN)	37.9 ± 0.1	57.5 ± 0.2	0.850 ± 0.001
	MDFS-CLASS (EN)	38.3 ± 0.2	57.6 ± 0.1	0.850 ± 0.000
	MDFS-NGRAM	38.4 ± 0.3	57.9 ± 0.2	0.852 ± 0.000
	MDFS-REGRESSION	38.0 ± 0.0	57.6 ± 0.0	0.849 ± 0.000
	MDFS-CLASS	37.9 ± 0.1	57.4 ± 0.0	0.849 ± 0.001

Table 15: Mean spBLEU, chrF++ and COMET scores for the En→Ar medical domain experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean. The threshold is millions of sentences.

Threshold	Method	spBLEU	chrF++	COMET
1.0	RANDOM	39.2 ± 0.2	58.3 ± 0.1	0.864 ± 0.001
	KEYWORD	41.8 ± 0.3	60.0 ± 0.1	0.878 ± 0.001
	MDFS-NGRAM (EN)	41.4 ± 0.0	59.6 ± 0.0	0.871 ± 0.001
	MDFS-REGRESSION (EN)	42.8 ± 0.1	60.6 ± 0.1	0.877 ± 0.000
	MDFS-CLASS (EN)	42.7 ± 0.3	60.6 ± 0.2	0.878 ± 0.000
	MDFS-NGRAM	41.4 ± 0.2	59.5 ± 0.1	0.866 ± 0.001
	MDFS-REGRESSION	39.9 ± 1.2	58.2 ± 0.5	0.837 ± 0.007
	MDFS-CLASS	42.0 ± 0.2	60.0 ± 0.1	0.873 ± 0.002
2.5	RANDOM	40.0 ± 0.2	58.9 ± 0.2	0.870 ± 0.001
	KEYWORD	42.5 ± 0.2	60.5 ± 0.1	0.880 ± 0.000
	MDFS-NGRAM (EN)	42.5 ± 0.1	60.5 ± 0.1	0.878 ± 0.001
	MDFS-REGRESSION (EN)	43.0 ± 0.2	60.7 ± 0.1	0.880 ± 0.001
	MDFS-CLASS (EN)	42.9 ± 0.3	60.7 ± 0.2	0.881 ± 0.001
	MDFS-NGRAM	42.4 ± 0.4	60.4 ± 0.3	0.880 ± 0.001
	MDFS-REGRESSION	42.1 ± 0.4	60.1 ± 0.2	0.874 ± 0.001
	MDFS-CLASS	42.7 ± 0.1	60.5 ± 0.1	0.875 ± 0.002
5.0	RANDOM	40.5 ± 0.2	59.1 ± 0.2	0.872 ± 0.000
	KEYWORD	42.1 ± 0.1	60.3 ± 0.1	0.878 ± 0.000
	MDFS-NGRAM (EN)	42.2 ± 0.2	60.3 ± 0.1	0.878 ± 0.000
	MDFS-REGRESSION (EN)	42.7 ± 0.2	60.6 ± 0.1	0.881 ± 0.000
	MDFS-CLASS (EN)	42.8 ± 0.2	60.6 ± 0.1	0.881 ± 0.000
	MDFS-NGRAM	42.5 ± 0.1	60.5 ± 0.1	0.880 ± 0.000
	MDFS-REGRESSION	42.8 ± 0.3	60.6 ± 0.2	0.881 ± 0.001
	MDFS-CLASS	42.3 ± 0.1	60.4 ± 0.1	0.881 ± 0.000
10.0	RANDOM	40.8 ± 0.1	59.4 ± 0.1	0.872 ± 0.001
	KEYWORD	41.8 ± 0.1	60.1 ± 0.1	0.877 ± 0.001
	MDFS-NGRAM (EN)	41.9 ± 0.2	60.0 ± 0.1	0.878 ± 0.001
	MDFS-REGRESSION (EN)	41.8 ± 0.1	60.0 ± 0.0	0.879 ± 0.000
	MDFS-CLASS (EN)	42.1 ± 0.0	60.3 ± 0.0	0.879 ± 0.000
	MDFS-NGRAM	41.6 ± 0.2	59.9 ± 0.1	0.880 ± 0.001
	MDFS-REGRESSION	42.2 ± 0.2	60.3 ± 0.1	0.879 ± 0.000
	MDFS-CLASS	42.1 ± 0.2	60.3 ± 0.1	0.880 ± 0.001

Table 16: Mean spBLEU, chrF++ and COMET scores for the En→Ro medical domain experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean. The threshold is millions of sentences.

We use sacrebleu[11] to calculate spBLEU scores, and the COMET scores use the default model[12].

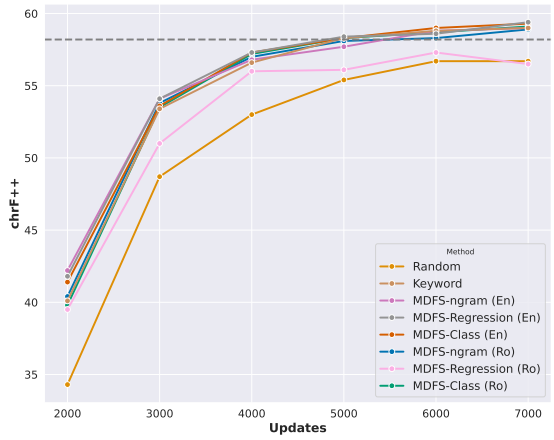

Figure 6: chrF++ scores for En→Ro evaluated on the test set with Beam Search plotted against updates in the range 2000-7000. The dashed line represents the maximal chrF++ achieved by the random baseline.

G.2 Learning Curve and LLM Labelled Test Set

Figure 6 shows the learning curves for the medical domain experiments on for En→Ro, evaluated on the test set using beam search with a beam size of 5 for updates 2000-7000. In addition to showing that filtering achieves the same performance as the RANDOM baseline after 5000 updates we also observe that the difference between filtering and the RANDOM decreases as a function of the updates.

We construct an additional test set for En→Ro from the 10,000 sentence test set (originally extracted from the training data) labelled with the LLM and used to evaluate the MDFS filtering models. We create the test set by taking all sentences that receive a score of 4 or higher from the LLM in the En→Ro direction and selecting the top 1,000 according to COMET. The chrF++ scores for these are given in Figure 7. The results evidence a larger improvement of the MDFS methods. The MDFS-NGRAM models achieve the highest chrF++ scores at a threshold of 1 million, and by extension are the best at translating sentences labelled as good medical examples by the LLM. This is in contradiction to the main results on the En→Ro test set which suggest that n-gram based approaches are weakest at low thresholds.

[11] https://github.com/mjpost/sacrebleu

[12] https://huggingface.co/Unbabel/wmt22-comet-da

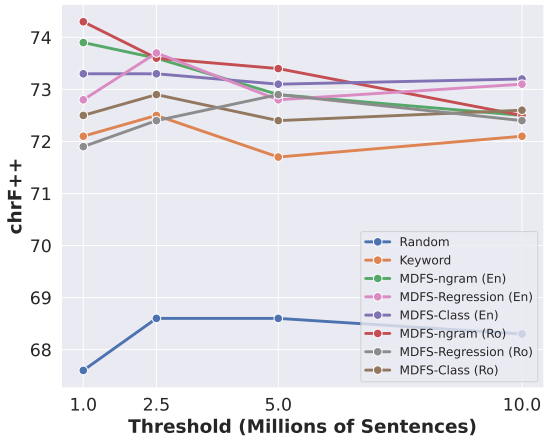

Figure 7: Mean chrF++ scores En→Ro reported on 1,000 best sentences according to COMET from the held-out test set labelled with `Llama-3.1-70B-Instruct` using three different random seeds. The errors are calculated using the Standard Error of the Mean.

H Domain-specific terminology evaluation details

We present the evaluation prompt in Figure 8. Following the findings of Qian et al. (2024), we include a chain of thought to the prompt to improve the LLM evaluation. The experiment was done using `gpt-4o-mini` as a judge.

The terminology evaluation experiment uses the 2.5 million threshold systems from the experiment depicted in Figure 7 and described in Section 5.2. As a representative of MDFS, we employ MDFS-CLASS (English).

```
Please find the medical word pairs in the source and
target language sentences. Refer to the above word
pairs to count the disambiguation accuracy in the
generated sentences of System A and System B.
Think step by step and produce a final score: 0 if
System A produced a better translation, 1 if it is a
tie, 2 if System B produced a better translation.

Source: "{source}"
Target: "{target}"
System A: "{system_a}"
System B: "{system_b}"
```

Figure 8: Template prompt used for medical terminology LLM-based evaluation.

I GPU Hours and Copilot Declaration

Code for this project was in parts written with the assistance of Copilot.

Labelling datasets with `Llama-3.1-70B-Instruct` was run on a single `A100-80GB` GPU. We labelled four datasets, each running taking around $\sim$ 70 hours.

Training MDFS models took $\sim$ 10 hours on either one `A100-40GB` GPU or two `RTX 3900` GPUs. Labelling the NMT data takes $\sim$ 24 hours, again run on either `A100-40GB` GPU or two `RTX 3900` GPUs. We train and predict twice for each language pair and task for a total of 8 runs.

NMT training and evaluation is run on either one `A100-40GB` GPU or one `RTX 3900`, with each run and evaluation taking $\sim$ 4 hours; we train 240 NMT models.