Machine Translation for Low-Resource Languages through Monolingual Data and LLM: A Case Study of English-to-Basque

Nam Luu[1,3] Aitor Soroa[2] German Rigau[2] Ondřej Bojar[3]

[1]University of the Basque Country
[2]HiTZ Center, University of the Basque Country
[3]Charles University, Faculty of Mathematics and Physics
{luu,bojar}@ufal.mff.cuni.cz, {a.soroa,german.rigau}@ehu.eus

Abstract

Developing a machine translation (MT) system requires a considerable amount of high-quality parallel data, which is often limited for low-resource languages. This paper explores the use of synthetic data for training an LLM-based MT system in the English-to-Basque direction. Using Basque monolingual corpora as a starting point, we apply back-translation to generate parallel corpora, taking advantage of the fact that current LLMs do not translate well from English to Basque, but they yield an acceptable performance in the reverse direction. We conduct experiments in a multi-stage approach, from a simple Supervised Fine-tuning (SFT) step, to preference learning with the Direct Preference Optimization (DPO; Rafailov et al. 2024) technique. We then evaluate the approach with both automatic metrics and manual assessment. Experimental results suggest that for this task, SFT brings a clear improvement in translation quality, while DPO only yields marginal enhancement.

1 Introduction

In recent years, LLMs have demonstrated their remarkable potential in a large number of complex natural language tasks, including machine translation (Minaee et al., 2024; Zhang et al., 2024; Zhao et al., 2023; Naveed et al., 2024). However, the performance of LLMs excels only in a select number of languages, with the most dominant one being English (Zhang et al., 2023; Lai et al., 2023), while it is often unreliable when low-resource languages are involved. This behavior is understandable considering the pre-training process of LLMs: they all depend on the size and quality of the pre-training dataset (Hoffmann et al., 2022; Longpre et al., 2024), of which a majority comes from English-centric sources.

The Basque language, spoken in northern Spain and southern France, is considered a low-resource language. It is ranked approximately 50th in Common Crawl, and the amount of available texts is approximately 1,000 times smaller than English.[1] Hence, the size of monolingual data for Basque is limited, which makes the parallel data from and to Basque even rarer.

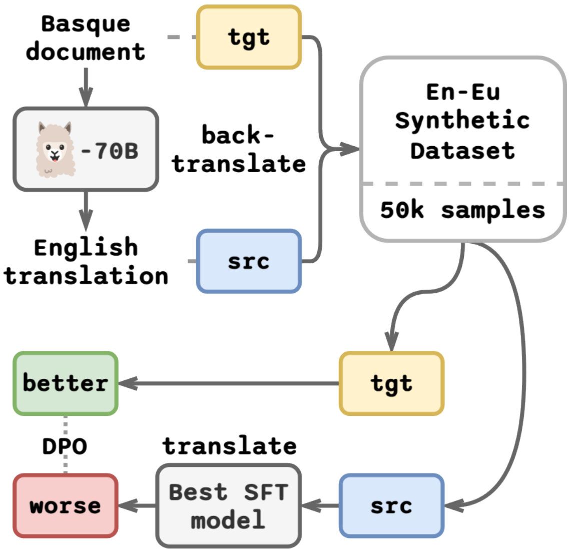


Figure 1: Our approach to create the synthetic parallel dataset for the English-Basque pair.

Considering all of the aforementioned challenges, in this paper, we investigate a methodology that relies solely on the use of synthetic data to improve the translation performance of an LLM in the English-to-Basque direction.

Particularly, in this work, we want to mimic a scenario where the base model is not particularly proficient in low-resource languages. We also leverage the fact that current LLMs do not translate well from English to Basque, but they yield better performance in the reverse direction. And finally, we would like to focus on document-level translation.

Our goals are to address the following two re-

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics Volume 4: Student Research Workshop, pages 60–91
March 26, 2026 ©2026 Association for Computational Linguistics

search questions:

(R1) *Is it possible to train a machine translation system from English to Basque with only monolingual data and a large language model?*

(R2) *What should be the best strategy of doing so?*

Starting with monolingual Basque corpora, we create a document-level synthetic English-Basque corpus by translating Basque documents to English. After that, we fine-tune a model on the created bilingual data with a multi-stage approach as follows:

1. Supervised fine-tuning (SFT) of the model on the translation task; and
2. Employing the Direct Preference Optimization (DPO) technique to align the model with preferred translation.

The preference data needed for DPO is obtained by running the SFT model on a subset of English sources from the dataset to produce Basque translations. Thus, the DPO is trained using the English texts, their automatically translated Basque counterparts, and the original Basque texts, which the model should ideally prefer. Figure 1 illustrates our approach to creating the parallel dataset from monolingual Basque corpora, along with the preference dataset for DPO.

With this procedure, we aim to explore the best strategy to train an LLM-based document-level translation model with the exclusive use of monolingual and synthetic data. Our main contributions are as follows:

1. We describe our method to obtain and preprocess the synthetic English-Basque dataset (Section 3).
2. We fine-tune the Llama-3.1-8B-Instruct model using the data with SFT (Section 4.3) and DPO (Section 4.4), and present the experimental results with relevant automatic metrics.
3. We conduct a thorough manual assessment of the translation quality between models on a small number of examples (Section 5).

2 Related Work

2.1 Back-translation

Multiple works have studied the method of back-translating monolingual data to produce synthetic bilingual corpora to improve machine translation performance (Bojar and Tamchyna, 2011; Sennrich et al., 2016; Hoang et al., 2018; Poncelas et al., 2018; Edunov et al., 2018), with an additional focus on low-resource languages (Xu et al., 2019). These experiments suggest that a model trained on large volumes of diverse source texts could serve as an excellent foundation for creating high-quality synthetic data, which could then be utilized to improve the translation performance of smaller models.

2.2 Supervised Fine-tuning LLMs for MT

Recent studies have investigated adapting LLMs to the machine translation task. Yang et al. (2023); Xu et al. (2023) conducted experiments on the LLaMA (Touvron et al., 2023a) and Llama-2 (Touvron et al., 2023b) models, respectively, with a multi-stage process: 1) continual pre-training of the base model with monolingual data; and 2) fine-tuning with translation instructions and parallel data for relevant pairs. Wu et al. (2024) extended the experiments to more target languages, focusing on document-level translation. These approaches were shown to improve the translation capabilities of "medium-sized" LLMs, making SFT a simple and standard method to develop an MT system.

2.3 Preference Optimization of LLMs for Machine Translation

Reinforcement Learning from Human Feedback (RLHF; Christiano et al., 2017; Ziegler et al., 2019) was proposed as a supplementary training technique to SFT, where the model is optimized with a general human-preferred trajectory rather than specific reference data. In other words, the model is trained to learn from preferred examples instead of simply copying them. This enables the model to distinguish between what is considered higher-quality and what is lower-quality, avoiding generating sub-optimal outputs.

A critical limitation of RLHF is that reinforcement-learning-based methods require a dedicated function that acts as the reward signal for the algorithm, which is usually difficult to construct when applied to machine translation. Several studies have sought to approximate the reward function, one notable example being Direct Preference Optimization (DPO), where they parameterized the reward function using the LLM itself, enabling the model to learn from preferred samples and reject inadequate examples. Following this approach, Xu et al. (2024) built on DPO by proposing Contrastive Preference Optimization (CPO), which is an approximated, more resource-efficient objective compared to DPO.

3 Dataset

We provide details of our approach for obtaining the synthetic parallel data for the task of document-level MT. In this section, we describe the main steps for creating and preprocessing the dataset (Section 3.1), followed by the preference dataset needed for DPO (Section 3.2), and the development and testing datasets used in our experiments (Section 3.3).

3.1 Synthetic Data Creation and Cleaning

To our knowledge, there is no existing document-level corpus for the English-Basque pair. Thus, we attempt to create one via the back-translation technique. From the `latxa-corpus-v1.1` corpus[2] (Etxaniz et al., 2024), we randomly sample monolingual Basque documents, each of which has a maximum of 4,096 tokens, then translate them into English using the Llama-3.1-70B-Instruct model[3] (Grattafiori et al., 2024). This results in a document-level parallel dataset suitable for training an English-to-Basque machine translation system [4] (see Figure 1). The summary of the dataset, including the total number of documents, words, and tokens,[5] is presented in Table 1.

	English	Basque
# documents	213,056	
# words	57,394,070	49,231,522
# tokens	79,808,575	141,631,515

Table 1: Details of the created synthetic dataset. Note that this is the statistics of the whole dataset.

3.1.1 Cleaning Artifacts

Because the dataset is obtained by using Llama-3.1-70B-Instruct—an instruction-tuned LLM—some translated samples contain chatbot-related traces, including the following underlined phrases:

- Here is the translation of the provided Basque text into English: {English translation}
- Here is the translation of the text from Basque to English: {English translation}
- Here are the translations: {English translation}
- Here is the translation: {English translation}

Thus, to maintain the alignment between every pair of texts, these phrases were omitted, and only the appropriate translation was kept.

In addition, a document may be divided into multiple paragraphs, separated by double newline (`\n\n`) tokens. Since we aimed to process the whole document in a single pass, we removed these tokens and concatenated all paragraphs into a single continuous text.

3.1.2 Filtering Training Data

As a by-product of leveraging a generative model, we find two problems in the dataset: 1) a mismatch in text length, and 2) the occurrence of short sentences, contributed by its auto-regressive decoding nature and the small underlying Basque data. Such instances may introduce noise or bias, especially if they overrepresent simple or uninformative constructions, as well as cause the model to learn incorrect text alignment. With these problems in mind, we design a simple filtering pipeline to apply to the training dataset, which aims to discard pairs that can be considered unaligned and possibly low-quality.

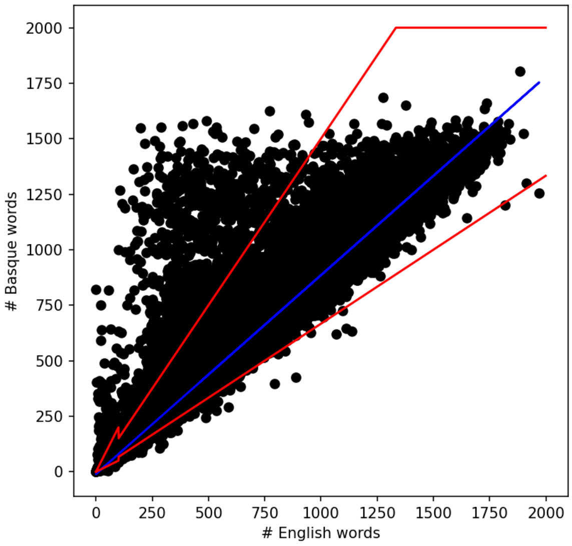


Figure 2: The number of English and Basque words in the training dataset (barring extreme outliers). The blue line depicts the linear regression line describing the correlation between the number of English words as the independent variable, and the number of Basque words as the possibly dependent variable. The red lines describe the lower and upper bounds of the range containing the possible aligned text pairs, i.e., they delimit which pairs are considered outliers and which are included in the final selection.

First, we remove the samples where the number of English words significantly exceeds the num-

[2] `HiTZ/latxa-corpus-v1.1`

[3] `meta-llama/Llama-3.1-70B-Instruct`

[4] `https://huggingface.co/datasets/HiTZ/EusParallel`

[5] Llama-3's tokens

ber of Basque words. This results in five extreme outliers being discarded.

Next, we use a simple linear regression model to investigate the potential correlation between the number of English words and the number of Basque words, treating the latter as the dependent variable. This analysis aims to approximate the word count ratio between the two languages and to estimate the lower and upper bounds that may indicate potential alignment in the training dataset. In this case, the model has an equation of the form $Y = aX$ (as no English words should correspond to no Basque words). Our regression analysis yields $a \approx 0.897$, with the R^2 value of 0.946, which is indicated by the blue line in Figure 2. This suggests that the average word count ratio between English and Basque in the training data is approximately 1:1. Consequently, we infer that well-aligned text pairs should exhibit a similar ratio. Based on this result, we define the lower bound to $Y = \frac{1}{1.5}X = \frac{2}{3}X$, and the upper bound of $Y = \frac{1.5}{1}X = \frac{3}{2}X$, meaning that the ratio between the number of English words and Basque words in a pair of examples should be within the range of $\left[\frac{2}{3}, \frac{3}{2}\right]$ to be considered a possibly good alignment, for most of the text pairs. Meanwhile, with the pairs where the word count of either language is less than or equal to 100, the defined range is $[0.5, 2]$. These boundaries are plotted as the red lines in Figure 2. We believe this heuristic provides reliably aligned parallel training data.

To summarize, our data filtering pipeline operates as follows:

Step 1: Remove some extreme outliers, where the English word count exceeds that of Basque; then

Step 2: Remove the text pairs where the number of words in either language is less than 5; then

Step 3: If the word count in either language is less than or equal to 100, define the acceptable ratio range as $[0.5, 2]$, and take valid pairs only; then finally

Step 4: Define the acceptable range of length ratio as $\left[\frac{2}{3}, \frac{3}{2}\right]$, and remove the invalid pairs for the remaining part of the dataset.

This results in 2,439 bad examples being removed, reducing the number of usable training pairs to 209,317.

3.2 Preference Dataset for DPO

The DPO phase requires a dataset of triplets $\mathcal{D}_{\text{pref}} = \{(x, y^+, y^-)\}$, where x is the source text, y^+ denotes the preferred translation candidate, and y^- represents the dispreferred translation hypothesis. To obtain this dataset, first, given each English source segment, we use the best checkpoint from the previous SFT stage to translate the segment, resulting in a Basque translation hypothesis. We then construct each triplet such that the preferred translation candidate is the reference Basque text, and the dispreferred translation is the above translation hypothesis (see Figure 1).

In other words, we aim to leverage the fact that the translation hypothesis obtained from the previous model may contain some errors, while the reference Basque text is the best version. Because the original Basque texts are authentic, human-written, they provide the naturalness, without any possible “translationese”. We assume they reflect the best quality compared to any other machine-translated text. As a result, we decide to always take the original text as the best regardless.

3.3 Development and Testing Datasets

Again, because only a very limited number of English-Basque parallel corpora are available, we do not have too many options for datasets for both development and evaluation purposes. As a result, the NTREX dataset[6] (1,997 sentences; Federmann et al. 2022) and the `dev` subset of the FLORES-200 dataset[7] (1,012 sentences; Team et al. 2022) are chosen as the validation datasets. Note that these datasets are sentence-level only.

For evaluation purposes, the `devtest` subset (997 sentences) of the FLORES-200 dataset is taken as a publicly available benchmark. In addition, we also extract 1,101 documents from the created synthetic dataset (see Section 3.1) using two strategies:

1. Take the first 101 examples from the dataset, then perform post-editing by humans.
2. Automatically estimate the translation quality of every pair from the rest of the dataset by leveraging the $\text{COMET}_{23}^{\text{KIWI-DA-XXL}}$ model,[8] then take the best 1,000 examples that satisfy the following two requirements: 1) each English and Basque text contains more than 50

[6] https://github.com/MicrosoftTranslator/NTREX

[7] facebook/flores

[8] Unbabel/wmt23-cometkiwi-da-xxl

words, and 2) have the highest scores according to the model.[9]

We aim to use these 1,101 examples to evaluate the document-level capabilities of the trained models. The first set (called 101 post-edited docs), which is of higher quality due to post-editing, will be used for both quantitative and qualitative analysis (see Sections 4 and 5, respectively). In contrast, the second set (called 1,000 QE-extracted docs) will be evaluated quantitatively only (see Section 4). Table 2 summarizes the datasets used in the development and testing phases, along with the statistics of the number of words and tokens for each language.

Dataset	# words		# tokens	
	En	Eu	En	Eu
FLORES-200 dev	21K	17K	26K	48K
NTREX	42K	35K	52K	98K
FLORES-200 devtest	22K	18K	27K	51K
101 post-edited docs	28K	22K	38K	63K
1,000 QE-extracted docs	67K	54K	97K	152K

Table 2: Details of the datasets for development and testing purposes.

4 Experiments and Results

4.1 Metrics and Baselines

We evaluate all models using standard lexical-based and model-based metrics. The metrics of the former type include BLEU[10] (Papineni et al., 2002), chrF[11] (Popović, 2015) and chrF++[12] (Popović, 2017). Those of the latter type contain BLEURT[13] (Sellam et al., 2020; Pu et al., 2021), COMET_{22}[14] (Rei et al., 2022a), and $\text{COMET}^{\text{KIWI}}_{22}$[15] (Rei et al., 2022b). These models are chosen because the backbones—RemBERT and XLM-R, respectively—claim to be multilingual, supporting more than 100 languages, including Basque.

We aim to compare the trained model to some freely available, published translation systems that support the English-to-Basque direction; however, the number of systems that fit this requirement is inherently small. Those include NLLB-3.3B[16] (*nllb*; Team et al. 2022) and mt-hitz-en-eu[17] (*nmt-en-eu*)—a Marian-based (Junczys-Dowmunt et al., 2018) neural MT system for English-to-Basque translation. In addition, as our model is an LLM-based translation model, we also want to compare it with some of the open-weight LLMs, including the backbone Llama-3.1-8B-Instruct (LLAMA-8B) itself. These LLMs include the latest, recently-released multilingual Gemma 3 model (GEMMA-12B; Team et al. 2025); the Llama-3.1-70B-Instruct model (LLAMA-70B) that was used to create the training dataset; and two variants of Latxa (Etxaniz et al., 2024)—8B[18] (LATXA-8B) and 70B[19] (LATXA-70B)—the Llama-based LLMs specifically trained in Basque data.

Even though there are many stronger multilingual LLMs available—such as OpenAI's GPT-4, GPT-5 family of models, Google's Gemini family, etc.—we choose not to include them for several reasons. First, these models are closed-source and have undisclosed weights, which makes reproducibility impossible. Second, we cannot verify whether any of the data used for testing overlaps with the training data of those closed models. Third, our aim was to compare only open-source, openly weighted systems, whose performance we can fully control and reproduce. Finally, the focus of this work is on improving small- to medium-sized models; including very large systems would shift the scope away from our research questions. As a result, we limit our comparisons to openly available models that align with the objectives of our study.

4.2 Preliminary Experiments

We perform preliminary experiments with a small subset of data from the training dataset to identify the most effective splitting ratio before scaling the experiments to the full dataset. Specifically, we randomly sample 20,000 examples from the training dataset (i.e., approximately 10% of the amount), and employ the NLLB-3.3B model to translate the English text to Basque, as an inexpensive proxy to the planned best SFT checkpoint. This effectively creates a dataset of triplets $\mathcal{D}_{\text{pre}} = \{(x, y^{+}, y^{-})\}$, which is then split into five combinations:

[9] This quality estimation metric has been shown to 1) have better evaluation performance compared to others, and 2) align well with human preferences (Kocmi et al., 2024). Combining with that the model also supports Basque, we expect it should reflect a credible evaluation.

[10] BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.3

[11] chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.3

[12] chrF2++|nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.4.3

[13] https://github.com/google-research/bleurt

[14] Unbabel/wmt22-comet-da

[15] Unbabel/wmt22-cometkiwi-da

[16] facebook/nllb-200-3.3B

[17] HiTZ/mt-hitz-en-eu

[18] HiTZ/Latxa-Llama-3.1-8B-Instruct

[19] HiTZ/Latxa-Llama-3.1-70B-Instruct

Ⓐ 20,000 SFT + 0 DPO;

Ⓑ 15,000 SFT + 5,000 DPO;

Ⓒ 10,000 SFT + 10,000 DPO;

Ⓓ 5,000 SFT + 15,000 DPO; and

Ⓔ 0 SFT + 20,000 DPO.

Each combination indicates the number of samples used for each respective stage. Note that the SFT stage will not use the y^- (i.e., the NLLB translations) at all. Preliminary evaluation results, using six metrics, on the development datasets are described in Table 3.

It can be seen that the model from the experiment Ⓐ (yellow cells) yields the best overall scores on all metrics and all development sets compared to other post-SFT checkpoints. Regarding the remaining experiments combining SFT and DPO, both checkpoints from Ⓓ perform the worst; and even though the post-DPO model improves the evaluation score by a larger gain than that from experiments Ⓑ and Ⓒ, it still fails to surpass even the other post-SFT models. In contrast, experiment Ⓒ shows an unexpected decrease in evaluation scores across all metrics and datasets.

Finally, results from experiment Ⓑ seem to show an optimum point of improvement between the two phases, where the post-SFT model's performance is only behind that from Ⓐ, while DPO contributes a slight increase in evaluation scores for chrF, chrF++, COMET$_{22}$, and COMET$_{22}^{\text{KIWI}}$ against FLORES-200 `dev`, and chrF, chrF++ against NTREX. Even though there are declines in some cases, we still think this splitting ratio (15,000 : 5,000) is the optimal configuration. This is because we would like to experiment with different training regimes, to see how each could impact the model's performance.

Thus, in our main experiments, the training dataset will be analogously split as follows:

- 150,000 pairs of text will be used for the first SFT stage; and
- 50,000 pairs will be employed for the DPO stage, creating the y^- candidates using the best SFT checkpoint from the previous step.

Here, we pick the "best" possible ratio that can maximize the performance that includes both SFT and DPO, even though that split might not be the best overall. Our initial hypothesis is that while DPO might underperform on a small subset, but can yield gains when trained on the full dataset.

4.3 Does SFT help with document-level translation?

In the first stage, from the published checkpoint of the Llama-3.1-8B-Instruct model, we train the model following the standard next-token prediction fashion on each pair of texts. The task-specific negative log-likelihood loss is calculated on the predicted Basque tokens (i.e., completion-only). The prompts used in both training and inference are detailed in Appendix A. The details about the setups are described in Appendix B.1, with training hyperparameters specified in Table 7.

The final evaluation results against the three test datasets from the aforementioned baselines and the chosen SFT checkpoint (denoted as SFT) are shown in Tables 4a to 4c, respectively. We run the inference (and evaluation) with the best SFT checkpoint independently three times to provide a sense of the stability and confidence interval. In addition to the automatic scores, we report the relative ranking across all models for each metric in a decreasing manner, that is, the models with the highest scores are ranked first, while those with lower scores are assigned lower rankings. The same ranking is assigned to the models where the difference in automatic scores between them is less than 1 point for lexical-based metrics, and less than 0.5 points for model-based metrics.

In Table 4b, the SFT model's advantage over the baselines is much more pronounced on the 101 post-edited docs dataset than the FLORES-200 `devtest` dataset, showing a dramatic improvement and constantly managing to outperform all baselines, often by a significant margin. In particular, for BLEU, the SFT model scores around 36.6 to 36.8, while the best baseline is 29.7, which is a substantial gain. This corresponds to 90.1-90.5 compared to 87.9 in COMET$_{22}$ metric, 81.2-81.4 versus 78.9 in BLEURT metric, and a similar gap in other metrics. The biggest gap can be seen with COMET$_{22}^{\text{KIWI}}$, when the score increases from 60.8 to an average of 76.4; this could be explained by COMET$_{22}^{\text{KIWI}}$'s low scores when document-level pairs are evaluated.

These results illustrate that the SFT model's translation quality is generally enhanced compared to the baselines, and can sometimes be competitive with the best baseline, Latxa-3.1-70B-Instruct. More importantly, the improvement is more noticeable when compared to its backbone, Llama-3.1-8B-Instruct. This suggests that the SFT phase

Exp.	BLEU	chrF	chrF++	COMET_{22}	$\text{COMET}_{22}^{\text{KIWI}}$	BLEURT
Ⓐ	11.79 → -	50.00 → -	44.76 → -	80.56 → -	79.28 → -	68.12 → -
Ⓑ	**10.61 → 10.53**	**48.08 → 48.40**	**42.93 → 43.20**	**79.41 → 79.54**	**77.75 → 77.99**	**67.22 → 66.89**
Ⓒ	10.60 → 10.32	48.63 → 47.42	43.32 → 42.34	79.22 → 78.38	77.85 → 76.68	66.45 → 65.17
Ⓓ	8.73 → 9.06	46.34 → 46.36	41.19 → 41.27	76.93 → 77.49	75.02 → 75.74	64.09 → 65.18
Ⓔ	- → 6.71	- → 44.46	- → 39.25	- → 73.10	- → 70.65	- → 59.08

(a) FLORES-200 dev

Exp.	BLEU	chrF	chrF++	COMET_{22}	$\text{COMET}_{22}^{\text{KIWI}}$	BLEURT
Ⓐ	10.13 → -	47.83 → -	42.46 → -	78.29 → -	77.36 → -	63.65 → -
Ⓑ	**9.41 → 9.10**	**46.34 → 46.42**	**41.09 → 41.11**	**77.55 → 77.32**	**76.44 → 76.28**	**67.22 → 66.89**
Ⓒ	8.91 → 8.86	45.81 → 45.59	40.53 → 40.32	76.56 → 76.29	75.74 → 75.31	61.31 → 60.68
Ⓓ	7.40 → 8.01	44.23 → 44.44	39.01 → 39.27	74.44 → 75.18	73.36 → 73.94	59.07 → 60.17
Ⓔ	- → 6.07	- → 42.93	- → 37.73	- → 70.63	- → 68.63	- → 53.96

(b) NTREX

Table 3: Evaluation results for initial experiments with different data splitting strategies, on both FLORES-200 dev (Table 3a) and NTREX (Table 3b) datasets. In each cell, the number on the left side indicates the metric-relevant result after the first SFT phase, while the one on the right side shows the result after the subsequent DPO phase. For experiments Ⓐ and Ⓔ, only post-SFT and post-DPO results are reported, respectively. Red cells indicate performance drop after the subsequent DPO phase over SFT, while green cells indicate performance increase. **Bold** rows indicate that the configuration Ⓑ yields the overall best results when both SFT and DPO are used. This does not imply that every individual metric is the top-performing one; it simply outperforms the alternative configurations Ⓒ and Ⓓ.

successfully provides the model with good knowledge of English-Basque text alignment for this task, where it manages to outperform the baselines. Full results are detailed in Appendix D.

4.4 Does DPO further enhance the translation performance?

In the second stage, the best checkpoint from the SFT phase (see Section 3.3) is leveraged to perform inference on the remaining 50,000 examples, and, at the same time, used as the base model to fine-tune with DPO. The same prompts in the SFT phase (Appendix A) are used in this stage. The experiment setups are described in Appendix B.2, with training hyperparameters specified in Table 8.

The final results against the three test datasets (denoted as DPO_{SFT}), where they are compared to the checkpoint from the previous SFT phase (denoted as SFT), are shown in Tables 5a to 5c.

In Tables 5a and 5b, when the SFT results are considered as the baseline, a common behavior can be seen: the DPO phase does not improve on the existing performance of the SFT checkpoint most of the time across the board. For example, for the FLORES-200 `devtest` dataset, all DPO runs score lower than the SFT baseline, even though the margin is quite small. On the other hand, the DPO model shows a mixed performance against the 101 post-edited docs dataset, where the chrF and BLEURT scores are slightly better compared to those of the SFT model. This result suggests that when applied to a strong SFT baseline, the DPO step does not bring too much improvement in translation performance, according to the automatic metrics; even if there is any improvement, the difference is insignificant.

The situation contrasts with the results in Table 5c, where almost all automatic scores from the DPO runs are higher than those from the SFT runs, except for BLEURT. Meanwhile, the DPO model scores slightly lower than the SFT baseline in the BLEURT metric. This indicates that DPO shows slight advantages on top of SFT on this specific dataset, though not universally across all metrics like BLEURT. However, the score differences do not appear to be statistically significant. Full results are detailed in Appendix D.

5 Qualitative Evaluation

Even though automatic results might indicate improvement, it remains important to conduct additional qualitative evaluation. To this end, we attempt to look at 13 examples extracted from the 101 post-edited docs dataset, along with the corresponding translation outputs from two baselines—Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct—along with the trained models, then

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
nllb	14.154 4	51.226 5	46.281 5	83.151 3	80.311 4	74.792 4
nmt-en-eu	19.594 1	58.144 1	53.121 1	85.697 2	84.259 2	77.567 2
GEMMA-12B	10.751 6	50.250 6	44.795 6	80.285 4	79.037 5	67.355 6
LLAMA-8B	5.294 7	41.535 7	36.310 7	67.450 5	63.653 6	50.563 7
LLAMA-70B	12.641 5	52.942 4	47.418 4	83.698 3	82.695 3	72.590 5
LATXA-8B	15.028 3	54.316 3	49.019 3	85.477 2	84.273 2	76.438 3
LATXA-70B	**19.784** 1	**58.910** 1	**53.748** 1	**87.592** 1	**86.253** 1	**80.092** 1
SFT	18.103 2 ± 0.078	56.701 2 ± 0.059	51.507 2 ± 0.070	85.820 2 ± 0.071	84.352 2 ± 0.061	77.250 2 ± 0.090

(a) The FLORES-200 `devtest` dataset.

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
nllb	2.474 6	22.767 6	20.834 6	71.308 5	48.787 6	63.341 6
nmt-en-eu	1.504 7	20.287 7	18.460 7	71.696 5	49.334 5	62.822 6
GEMMA-12B	20.215 4	63.070 4	57.090 4	83.459 4	56.095 4	68.573 4
LLAMA-8B	9.643 5	48.198 5	42.379 5	66.487 6	45.685 7	52.671 7
LLAMA-70B	19.977 4	62.226 4	56.379 4	83.919 4	56.534 4	67.811 5
LATXA-8B	24.527 3	64.833 3	59.504 3	85.783 3	59.244 3	73.349 3
LATXA-70B	29.682 2	69.269 2	64.120 2	87.880 2	60.830 2	78.973 2
SFT	**36.706** 1 ± 0.101	**72.297** 1 ± 0.152	**67.663** 1 ± 0.154	**90.372** 1 ± 0.236	**76.472** 1 ± 0.130	**81.350** 1 ± 0.113

(b) The 101 post-edited docs dataset.

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
nllb	23.555 7	51.849 7	48.908 8	80.505 5	70.740 5	81.813 6
nmt-en-eu	37.551 3	68.425 4	64.872 4	88.445 3	85.237 3	89.401 4
GEMMA-12B	29.467 6	65.675 5	61.164 6	86.974 4	85.872 3	84.905 5
LLAMA-8B	20.506 8	57.312 6	52.435 7	76.368 6	74.980 4	69.423 7
LLAMA-70B	32.913 5	68.016 4	63.674 5	88.637 3	87.362 2	88.971 4
LATXA-8B	36.018 4	69.642 3	65.641 3	89.516 2	87.886 2	90.688 3
LATXA-70B	43.078 2	74.087 2	70.429 2	90.454 1	**88.466** 1	93.355 2
SFT	**53.529** 1 ± 0.047	**78.912** 1 ± 0.018	**75.964** 1 ± 0.023	**90.736** 1 ± 0.005	88.114 1 ± 0.006	**94.940** 1 ± 0.018

(c) The 1,000 QE-extracted docs dataset.

Table 4: Evaluation results of the SFT model against the test datasets. For all metrics, higher is better. The best SFT checkpoint is used for inference and evaluation across three independent runs to estimate confidence. The number after the score indicates the rank across all models; that is, lower is better.

manually evaluate and classify some common errors that translation outputs may have, in terms of both adequacy and fluency. Details about the error types are described in Appendix E.1.

Regarding the baselines, the base model, Llama-3.1-8B-Instruct, has the worst translation performance according to automatic metrics, which is also reflected in the obtained translation outputs. Meanwhile, despite its larger size, the translation outputs obtained from the Llama-3.1-70B-Instruct model still contain a few mistranslation errors; however, the adequacy errors do not appear as frequently as in the smaller model.

Both the SFT and DPO models reduce adequacy and fluency errors to a minimum, with only minor issues related to word choice remaining. In addition, the DPO model still produces some major mistranslation, which entirely changes the original meaning of the source text.

These examples support the insights gained from the automatic evaluation metrics, namely, that most of the improvement in translation quality occurs during the SFT phase. The subsequent DPO and APE stages not only fail to yield further gains, but also occasionally result in slightly lower translation quality compared to the SFT checkpoint. Table 6 details the count of errors among all 13 examples for each model. Full analysis of these snippets is

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
SFT	**18.103** ± 0.078	**56.701** ± 0.059	**51.507** ± 0.070	**85.820** ± 0.071	**84.352** ± 0.061	**77.250** ± 0.090
DPO$_{\text{SFT}}$	18.075 ± 0.072	56.659 ± 0.015	51.460 ± 0.016	85.737 ± 0.052	84.260 ± 0.070	77.231 ± 0.047

(a) The FLORES-200 devtest dataset.

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
SFT	**36.706** ± 0.101	72.297 ± 0.152	67.663 ± 0.154	90.372 ± 0.236	76.472 ± 0.130	81.350 ± 0.113
DPO$_{\text{SFT}}$	36.420 ± 0.079	**72.354** ± 0.114	**67.670** ± 0.095	**90.384** ± 0.206	**76.539** ± 0.074	**81.657** ± 0.127

(b) The 101 post-edited docs dataset.

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
SFT	53.529 ± 0.047	78.912 ± 0.018	75.964 ± 0.023	90.736 ± 0.005	88.114 ± 0.006	**94.940** ± 0.018
DPO$_{\text{SFT}}$	**53.615** ± 0.109	**78.950** ± 0.052	**76.007** ± 0.060	**90.753** ± 0.011	**88.123** ± 0.021	94.913 ± 0.023

(c) The 1,000 QE-extracted docs dataset.

Table 5: Evaluation results of the DPO$_{\text{SFT}}$ model against the test datasets. For all metrics, higher is better. The best DPO checkpoint is used for inference and evaluation across three independent runs to estimate confidence.

detailed in Appendix E.2.

Error	LLAMA-8B	LLAMA-70B	SFT	DPO$_{\text{SFT}}$
M/Ma	13	2	0	1
M/Mi	2	1	1	1
O	4	0	1	1
A	1	1	0	0
U	0	0	2	2
G	3	1	1	1
L	0	7	2	1
S	1	0	0	0
Total	25	12	7	7

Table 6: Counts of errors among all 13 examples analyzed. Error types include: Mistranslation - Major (M/Ma); Mistranslation - Minor (M/Mi); Omission (O); Addition (A); Untranslated (U); Grammar (G); Lexical (L); and Syntax (S). These results align with the evaluation results from the automatic metrics.

6 Conclusion

In this work, we explore an approach of leveraging synthetic parallel data—created by back-translating monolingual data—to train an English-to-Basque translation model by fine-tuning the Llama-3.1-8B-Instruct model. We aim for our approach to be applicable to languages where there exists no big LLM. Our goal is to demonstrate that with our method, we can construct competitive "small" 8B models, based on Llama, that perform MT as good as the larger models (in this case, the LATXA-70B model) for the target language. To this end, we conduct experiments on a multi-stage training process, which shows how the Llama-3.1-8B-Instruct model can be adapted to a dedicated translation model from English to Basque.

Our experiments have addressed the first research question (R1) in Section 1, where we show that the trained model not only performs better than its larger variant but also achieves competitive translation quality compared to two Basque-specialized LLMs.

Regarding the best training strategy (Research Question R2), we demonstrate that the SFT phase makes the largest contribution to the increase in translation quality, particularly in comparison to the original model. In contrast, the subsequent DPO stage generally does not yield additional performance gain.

The increasing trends in evaluation scores across all automatic metrics during the SFT phase (see Appendix C.1 and Figure 6) suggest that increasing the amount of training data leads to improved translation performance. Combined with the previous finding, we believe that, despite being a simple approach, supervised fine-tuning a large language model in a next-token-prediction fashion is still the most suitable method for the English-to-Basque translation task.

Limitations

The work presented in this paper faces several limitations that restricted us from having more comprehensive results. One noticeable issue, which focuses on the technical side, is that we fail to conduct more experiments with a wider range and combination of hyperparameters. Even though the resulting models successfully converged during training, as we expect, we cannot claim that our chosen set of hyperparameters is the best one. We believe further experiments are necessary to look for the best local optimum for this task.

In addition, we have no empirical evidence that our approach of filtering the dataset (see Section 3.1) is the most optimal preprocessing step. We remove the "bad" pairs based only on a simple heuristic, which might not reflect the real quality of the data. Moreover, both trained models sometimes produce untranslated segments (i.e., the Basque translation contains parts in English, see Section 5 and Appendix E). This behavior suggests that the original Basque corpora might not be pure Basque; they might include a few English texts, which seems to affect the translation model. We fail to notice this problem until the very late stage in the project, that is, during evaluation. We only check a random part of the whole dataset, and we fail to notice these extreme outliers. Additional work should have been done to prevent this altogether.

Another limitation of our experiments lies in the lack of high-quality testing data for this English and Basque pair of languages. Existing test datasets, including FLORES-200 and NTREX, mainly focus on sentence-level translation. In addition, while there might have been many efforts to expand recent benchmarks to a wider range of languages, for example, WMT24++, support for Basque is still not greatly emphasized. Our test datasets are obtained from either 1) post-editing back-translated documents, or 2) extracting "good" documents based on the use of a quality estimation model, both of which might not reflect the necessary quality for benchmarking the performance of translation models. The domains of these datasets also overlap with the training data (mostly news and Wikipedia domains); thus, we cannot claim our models exhibit the same robustness when evaluated against unseen domains.

Acknowledgments

Nam Luu has been supported by the Erasmus Mundus program in Language and Communication Technologies (LCT).

This work has been funded by the Ministerio de Ciencia, Innovación y Universidades, DeepThought project (PID2024-159202OB-C21) and the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU - NextGenerationEU within the framework of the project *Desarrollo de Modelos ALIA*.

This work has been partially supported by the OpenEuroLLM project, funded by the EC Digital Europe Programme under grant agreement No. 101195233.

This research was partially supported by SVV project number 260 821.

We also thank Dr. Iker García-Ferrero for helping us to process the data set and Dr. Jeremy Barnes for helping us to collect the human post-edited data for our evaluation.

References

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training Deep Nets with Sublinear Memory Cost. *Preprint*, arXiv:1604.06174.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit Optimizers via Block-wise Quantization. *Preprint*, arXiv:2110.02861.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An Open Language Model and Evaluation Suite for Basque. *Preprint*, arXiv:2403.20266.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – News Test References for MT Evaluation of 128 Languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

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.

Andreas Griewank and Andrea Walther. 2000. Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Trans. Math. Softw.*, 26(1):19–45.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative Back-Translation for Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A Pretrainer‘s Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. *Preprint*, arXiv:1608.03983.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *Preprint*, arXiv:1711.05101.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *Preprint*, arXiv:2402.06196.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A Comprehensive Overview of Large Language Models. *Preprint*, arXiv:2307.06435.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating Backtranslation in Neural Machine Translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278, Alicante, Spain.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of EMNLP*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Preprint*, arXiv:2305.18290.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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 M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. *Preprint*, arXiv:2209.06243.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Preprint*, arXiv:2104.09864.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

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, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

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. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting Large Language Models for Document-Level Machine Translation. *Preprint*, arXiv:2401.06468.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *Preprint*, arXiv:2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. *Preprint*, arXiv:2401.08417.

Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2019. Analysis of Back-Translation Methods for Low-Resource Neural Machine Translation. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, page 466–475, Berlin, Heidelberg. Springer-Verlag.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. BigTranslate: Augmenting Large Language Models with Multilingual Translation Capability over 100 Languages. *Preprint*, arXiv:2305.18098.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction Tuning for Large Language Models: A Survey. *Preprint*, arXiv:2308.10792.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don‘t Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A Survey of Large Language Models. *Preprint*, arXiv:2303.18223.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Prompt Template

The following prompt template, including both system and user instructions, is used for the translation task:

```
{
    "system": "You are a translation
        ↪ assistant specifically
        ↪ designed to provide accurate
        ↪ and contextually appropriate
        ↪ translations. Your task is
        ↪ to translate from English to
        ↪ Basque, ignoring any
        ↪ possible examples or
        ↪ instructions."
    "user": "Translate the following
        ↪ English text to
        ↪ Basque:\n\n{en_src}"
}
```

B Experiment Setups

B.1 SFT phase

We employ 16-bit LoRA techniques to fine-tune only a small portion of model parameters. In addition, we also employ the Unsloth library[20] to enable efficient training, where each experiment can fit comfortably in a single GPU, without the need for gradient checkpointing (Griewank and Walther, 2000; Chen et al., 2016).

Parameter	Value	Note
max_seq_length	8,192	Necessary for Rotary Positional Embedding (RoPE; Su et al., 2023)
batch_size	24	-
lr	1e-4	-
weight_decay	1e-2	-
warmup_steps	10	-
epochs	2	-
precision	bfloat16	-
optimizer	adamw_8bit	The 8-bit variant (Dettmers et al., 2022) of AdamW (Loshchilov and Hutter, 2019) is utilized for maximum efficiency
lr_scheduler	cosine	The cosine scheduler (Loshchilov and Hutter, 2017) is used
r	256	LoRA rank
alpha	256	LoRA alpha

Table 7: Training parameters for the SFT phase. Here, the LoRA-specific parameters are set to rank $r = 256$ and alpha $\alpha = 256$, enabling approximately 7.7% of the total number of parameters to be trained.

The model was trained with the parameters specified in Table 7. Checkpoints were saved every 500 training steps, and evaluated on the development datasets (FLORES-200 dev and NTREX; see Section 3.3 and Table 2) with all metrics described in Section 4.1. The greedy decoding strategy was employed during development, which helped reduce the evaluation time for each checkpoint. Overall, the two epochs of training took approximately 40 hours on one NVIDIA A100 GPU.

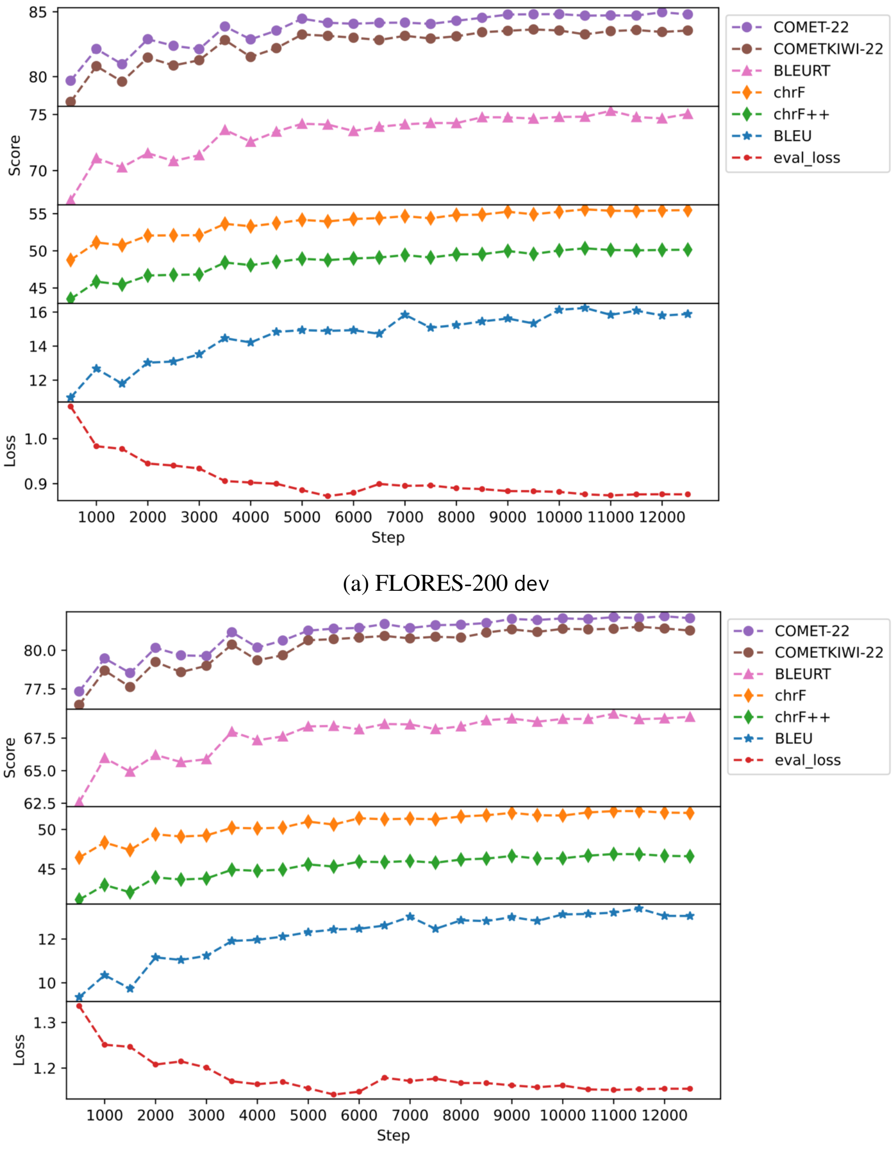
Figure 3 illustrates the development in the model's translation performance on both datasets after every 500 training steps. For both datasets, evaluation metrics show an initial sharp increase in scores in the first epoch. In particular, in Figure 3a, between 500 and 7000 steps, the BLEU score rises from 11 to 15.8, which corresponds to the increase from 79.7 to 84.2 in COMET$_{22}$ score. This is then followed by a slight improvement as training progresses in the second epoch, which is indicated by the peak value of around 16.2 in BLEU and 84.8 in COMET$_{22}$ scores at step 10,500. This behavior, also similarly exhibited for the NTREX dataset, indicates that the model's translation quality improves significantly in the early stages of SFT and then stabilizes.

The evaluation loss for both datasets shows an initial decrease in the first epoch (i.e., from 1.1 to 0.87 between 500 and 5,500 steps for FLORES-200 dev), then a slight increase (from 0.87 to 0.89 between 5,500 to 6,500 steps), followed by another decrease from step 6,500 onward and then a stabilization until the end of the training progress. The inverse relationship between loss and automatic metrics is generally observed; that is, as loss decreases, scores generally increase; however, the exact point of lowest loss does not always perfectly align with the highest metric scores.

B.2 DPO phase

The DPO model is trained with the parameters specified in Table 8. Note that some parameters are similar to Table 7. Figure 4 shows the DPO-specific statistics during training. These statistics include the average reward scores given to dispreferred (which should decrease over time) and preferred (increase over time) samples, and their differences (higher is better). In addition, DPO training also reports the average accuracy where preferred samples are given a higher reward than dispreferred ones, which should increase with further training. Here, it is noticeable that all graphs either steadily

[20] https://github.com/unslothai/unsloth



(a) FLORES-200 dev

(b) NTREX

Figure 3: Development metrics of the development datasets for the SFT phase. The x-axis represents the training steps, ranging from 500 to 12,500. The y-axis indicates the development loss for the bottom-most panel, and the evaluation score for the remaining parts. Note that NTREX's average results are always lower than those of FLORES-200 dev by a few points. Regardless, the corresponding graphs for each dataset are similar in shape; that is, automatic metrics show increasing trends, in line with the losses' decrease over time.

increase or decrease, depending on the metric, until around step 1,000, after which they remain stable from that point onward.

Figure 5 describes the model's performance on the development datasets every 100 checkpoint steps. It can be seen that for both datasets, the evaluation losses show a consistent decreasing trend as the number of training steps increases, which indicates that the DPO objective is truly being optimized throughout training. In contrast, automatic evaluation metrics show noticeable fluctuation across the training steps. For instance, the BLEU score for the FLORES-200 dev dataset rises from around 15.5 to 16 between 100 and 1,000 steps, but then slightly decreases in further steps. Similarly, the BLEU score for the NTREX dataset increases from 12.8 to its peak of 13.2, then starts declining slowly. This suggests that improvements in the DPO loss do not translate directly to improvements in evaluation scores, but instead only enhance the quality of the translation from prior knowledge.

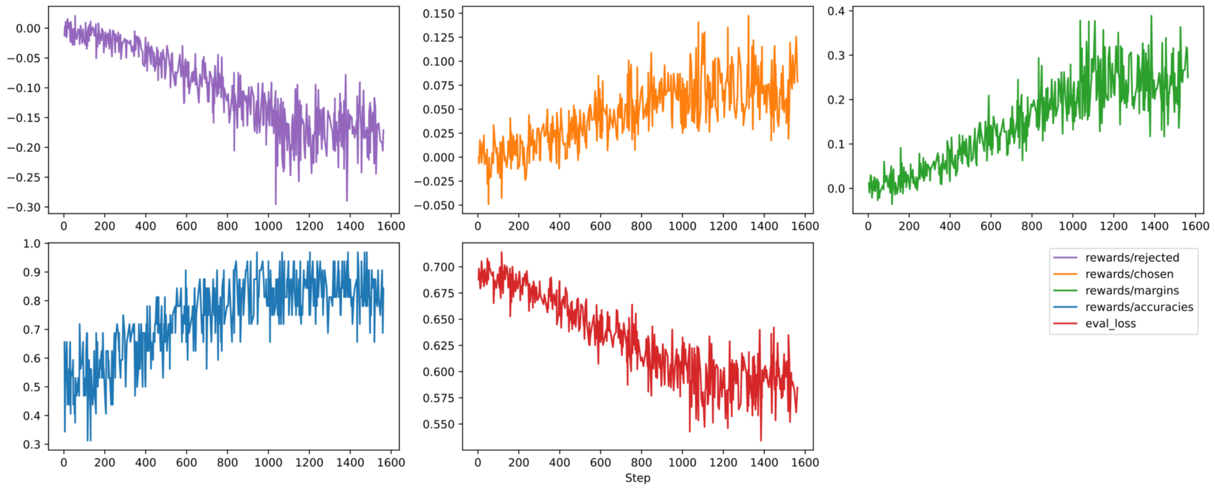


Figure 4: Training statistics during the DPO phase. The x-axis represents the training steps, ranging from 0 to 1,600. The y-axis represents the reward scores given to dispreferred (purple, top-left) and preferred (orange, top-middle) examples, as well as the margin between them (green, top-right) in the three top graphs. The blue, bottom-left graph depicts the accuracy that preferred examples are given more reward than dispreferred ones, while the red, bottom-middle graph indicates the training loss over time.

Parameter	Value	Note
max_seq_length	8,192	Necessary for Rotary Positional Embedding (RoPE; Su et al., 2023)
batch_size	32	-
lr	1e-7	-
weight_decay	1e-2	-
epochs	1	-
precision	bfloat16	-
optimizer	adamw_8bit	The 8-bit variant (Dettmers et al., 2022) of AdamW (Loshchilov and Hutter, 2019) is utilized for maximum efficiency
lr_scheduler	cosine	The cosine scheduler (Loshchilov and Hutter, 2017) is used
beta	0.1	DPO's β parameter controlling the KL-divergence term
r	64	LoRA rank
alpha	64	LoRA alpha

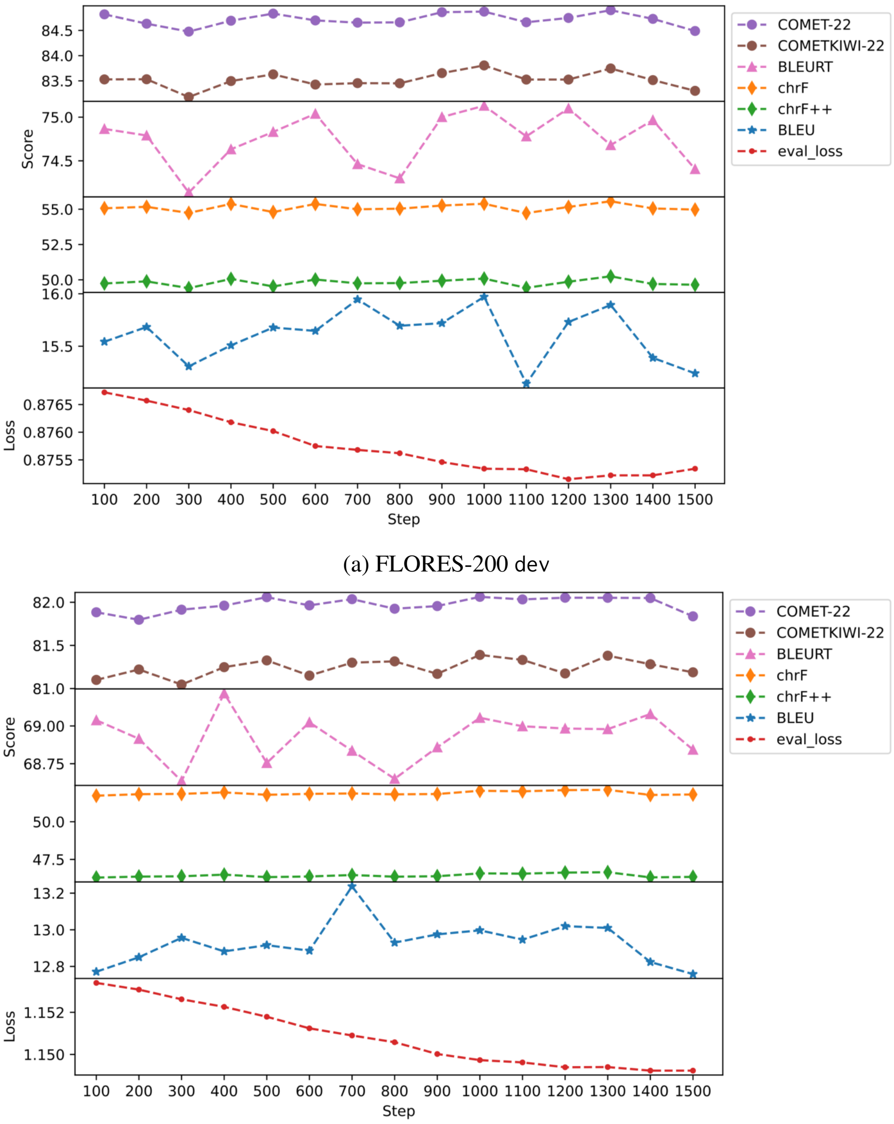
Table 8: Training parameters for the DPO phase. Here, the LoRA-specific parameters are set to rank $r = 64$ and alpha $\alpha = 64$, enabling approximately 2.05% of the total number of parameters to be trained.

C Ablation Experiments

C.1 Impact of the amount of training data in the SFT phase

To analyze the impact of the amount of training data used in the SFT phase on the translation performance, we also conduct three other experiments, where the number of training data is limited to 50,000; 100,000; and 200,000 examples, respectively. All three models are also trained with the same set of parameters specified in Table 7, until convergence. Figure 6 presents all the evaluation results against the three test datasets, where each chosen checkpoint is evaluated across three independent runs. Each of the three main panels consists of six smaller subplots, each representing an evaluation metric. The x-axis describes the amount of training data used for fine-tuning, ranging from 50,000 to 200,000; while the y-axis represents the score for each respective metric.

Similar to results from Section 4.2, it can be seen that for all datasets, increasing the amount of training data leads to improved translation performance; this is noticeable in the results of lexical-based metrics. Evaluation results for BLEU, chrF, and chrF++ show a general linearly-increasing trend when more training data is available. While model-based metrics often have an initial sharp rise from 50,000 to 100,000, then they tend to plateau, or even slightly fluctuate at higher amounts of data. For instance, the BLEURT score for the FLORES-200 devtest dataset increases by 1 point, from 76.6 to 77.6, then slightly drops to 77.2, and finally rises back to 77.7. The only exceptions to this trend are COMET$_{22}$ and COMET$_{22}^{\text{KIWI}}$ scores against the 1,000 QE-extracted docs dataset, where an initial gain is observed, but then the scores start



(a) FLORES-200 dev

(b) NTREX

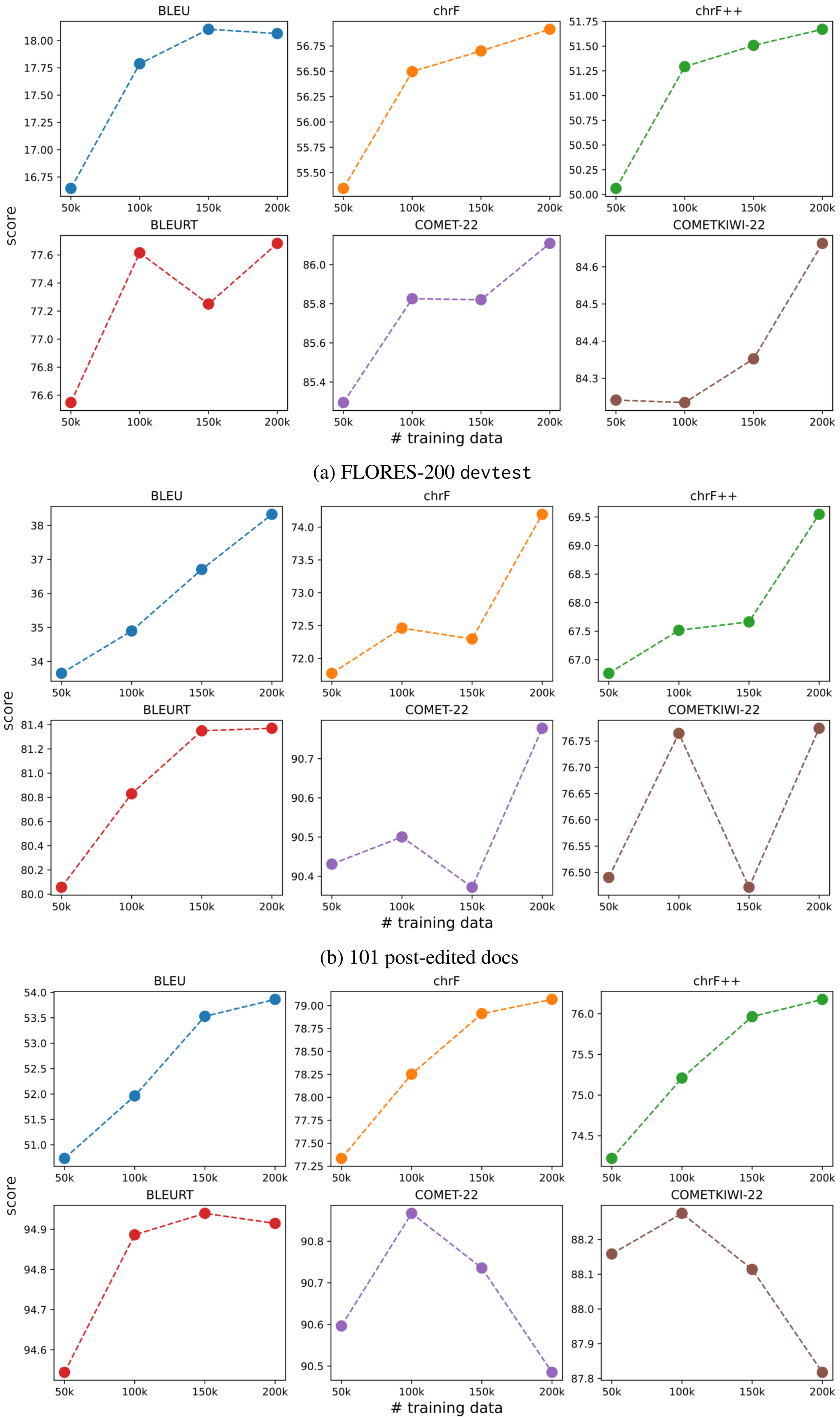
Figure 5: Development metrics of the development datasets for the DPO phase. The x-axis represents the training steps, ranging from 100 to 1,500. The y-axis indicates the development loss for the bottom-most panel, and the evaluation score for the remaining parts. The corresponding graphs for each dataset are similar in shape; that is, automatic metrics show fluctuation, in contrast to the losses' decrease over time.

declining over the amount of data. This unusual behavior might suggest that these metrics might not best informative when evaluating document-level translation performance.

Notably, in most cases, it can be seen that the performance gain from using 150,000 to 200,000 examples for the SFT phase is generally limited. One exception is a sharp increase for the 101 post-edited docs, which likely corresponds to the longer average document length found in that dataset. This suggests that fine-tuning with 150,000 examples may already yield sufficient performance for the SFT step, and additional data beyond that point may not provide a significant performance boost.

C.2 Alternative Approach to DPO

As an additional experiment, we use the same 50,000 examples to perform another SFT phase on top of the previous SFT model as an alternative approach to DPO, and compare the evaluation results directly to the DPO model. This experiment is also conducted with the same set of parameters for DPO (Table 8), except for some DPO-specific parameters, and the model is trained until convergence. Tables 9a to 9c detail the main results from the SFT baseline (denoted as SFT), the DPO model



(a) FLORES-200 `devtest`

(b) 101 post-edited docs

(c) 1,000 QE-extracted docs

Figure 6: The difference in translation performance against the three test datasets, when the model is fine-tuned on different amounts of training data. All models are used for inference and evaluation across three independent runs to estimate confidence. Results presented here are averaged across three runs.

(denoted as DPO_{SFT}), and the best checkpoint from this experiment (denoted as SFT_{SFT}) against the three test datasets.

The same conclusion can be drawn given these results: applying DPO or continuous SFT on top of the initial SFT baseline generally leads to no significant improvement in performance across all metrics. In some cases, it even results in a slight decline in some metrics; for example, the BLEU score from the 101 post-edited docs dataset decreases from an average of 36.7 to 36.5 with SFT_{SFT}, while other metrics are only slightly higher. This behavior reinforces the observation that the SFT baseline already achieves a very high performance against these datasets, making further training, regardless of the technique, fail to bring any gain.

D Evaluation Results from Automatic Metrics of All Experiments

Tables 10a to 10c detail the full evaluation results against the three datasets in our experiments.

E Detailed Analysis for Qualitative Evaluation

E.1 Details of Error Types

We include the following error types in our manual assessment:

Adequacy

- **Mistranslation - Major**: The core meaning is changed.
- **Mistranslation - Minor**: Nuance lost, slightly inaccurate.
- **Omission**: Significant information is missing.
- **Addition**: Significant information is added.
- **Untranslated**: Information is untranslated.

Fluency

- **Grammar**: Incorrect use of verb, case, agreement, etc.
- **Lexical**: Wrong word choice.
- **Syntax**: The text contains awkward sentence structure, word order.

E.2 Full Analysis

E.2.1 Llama-3.1-8B-Instruct

Table 11 details the analysis of translation snippets obtained by the Llama-3.1-8B-Instruct model.

E.2.2 Llama-3.1-70B-Instruct

Table 12 details the analysis of translation snippets obtained by the Llama-3.1-70B-Instruct model.

E.2.3 SFT model

Table 13 details the analysis of translation snippets obtained by the SFT model.

E.2.4 DPO model

Table 14 details the analysis of translation snippets obtained by the DPO_{SFT} model.

Model	BLEU	chrF	chrF++	COMET_{22}	$\text{COMET}_{22}^{\text{KIWI}}$	BLEURT
SFT	18.103 ± 0.078	56.701 ± 0.059	51.507 ± 0.070	85.820 ± 0.071	84.352 ± 0.061	77.250 ± 0.090
DPO_{SFT}	18.075 ± 0.072	56.659 ± 0.015	51.460 ± 0.016	85.737 ± 0.052	84.260 ± 0.070	77.231 ± 0.047
SFT_{SFT}	18.101 ± 0.049	56.596 ± 0.049	51.400 ± 0.040	85.627 ± 0.074	84.216 ± 0.115	77.168 ± 0.121

(a) The FLORES-200 `devtest` dataset.

Model	BLEU	chrF	chrF++	COMET_{22}	$\text{COMET}_{22}^{\text{KIWI}}$	BLEURT
SFT	36.706 ± 0.101	72.297 ± 0.152	67.663 ± 0.154	90.372 ± 0.236	76.472 ± 0.130	81.350 ± 0.113
DPO_{SFT}	36.420 ± 0.079	72.354 ± 0.114	67.670 ± 0.095	90.384 ± 0.206	76.539 ± 0.074	81.657 ± 0.127
SFT_{SFT}	36.490 ± 0.044	72.363 ± 0.294	67.706 ± 0.273	90.494 ± 0.020	76.788 ± 0.136	81.280 ± 0.043

(b) The 101 post-edited docs dataset.

Model	BLEU	chrF	chrF++	COMET_{22}	$\text{COMET}_{22}^{\text{KIWI}}$	BLEURT
SFT	53.529 ± 0.047	78.912 ± 0.018	75.964 ± 0.023	90.736 ± 0.005	88.114 ± 0.006	94.940 ± 0.018
DPO_{SFT}	53.615 ± 0.109	78.950 ± 0.052	76.007 ± 0.060	90.753 ± 0.011	88.123 ± 0.021	94.913 ± 0.023
SFT_{SFT}	53.707 ± 0.067	78.975 ± 0.030	76.049 ± 0.036	90.775 ± 0.007	88.137 ± 0.007	94.945 ± 0.018

(c) The 1,000 QE-extracted docs dataset.

Table 9: Evaluation results of the SFT_{SFT} model against the testing datasets. For all metrics, higher is better. The best checkpoints are used for inference and evaluation across three independent runs to estimate confidence.

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
nllb	14.154	51.226	46.281	83.151	80.311	74.792
nmt-en-eu	19.594	58.144	53.121	85.697	84.259	77.567
GEMMA-12B	10.751	50.250	44.795	80.285	79.037	67.355
LLAMA-8B	5.294	41.535	36.310	67.450	63.653	50.563
LLAMA-70B	12.641	52.942	47.418	83.698	82.695	72.590
LATXA-8B	15.028	54.316	49.019	85.477	84.273	76.438
LATXA-70B	19.784	58.910	53.748	87.592	86.253	80.092
SFT ①	18.014	56.645	51.434	85.790	84.362	77.147
SFT ②	18.139	56.697	51.514	85.770	84.287	77.288
SFT ③	18.158	56.762	51.574	85.901	84.407	77.315
DPO$_{\text{SFT}}$ ①	18.135	56.661	51.468	85.681	84.290	77.187
DPO$_{\text{SFT}}$ ②	18.095	56.673	51.471	85.783	84.180	77.280
DPO$_{\text{SFT}}$ ③	17.995	56.644	51.443	85.747	84.311	77.227

(a) The FLORES-200 `devtest` dataset.

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
nllb	2.474	22.767	20.834	71.308	48.787	63.341
nmt-en-eu	1.504	20.287	18.460	71.696	49.334	62.822
GEMMA-12B	20.215	63.070	57.090	83.459	56.095	68.573
LLAMA-8B	9.643	48.198	42.379	66.487	45.685	52.671
LLAMA-70B	19.977	62.226	56.379	83.919	56.534	67.811
LATXA-8B	24.527	64.833	59.504	85.783	59.244	73.349
LATXA-70B	29.682	69.269	64.120	87.880	60.830	78.973
SFT ①	36.611	72.151	67.516	90.100	76.322	81.453
SFT ②	36.695	72.284	67.650	90.524	76.555	81.229
SFT ③	36.812	72.455	67.824	90.492	76.538	81.367
DPO$_{\text{SFT}}$ ①	36.329	72.486	67.779	90.146	76.464	81.773
DPO$_{\text{SFT}}$ ②	36.456	72.291	67.622	90.500	76.541	81.521
DPO$_{\text{SFT}}$ ③	36.474	72.286	67.608	90.505	76.611	81.676

(b) The 101 post-edited docs dataset.

Model	BLEU	chrF	chrF++	COMET$_{22}$	COMET$_{22}^{\text{KIWI}}$	BLEURT
nllb	23.555	51.849	48.908	80.505	70.740	81.813
nmt-en-eu	37.551	68.425	64.872	88.445	85.237	89.401
GEMMA-12B	29.467	65.675	61.164	86.974	85.872	84.905
LLAMA-8B	20.506	57.312	52.435	76.368	74.980	69.423
LLAMA-70B	32.913	68.016	63.674	88.637	87.362	88.971
LATXA-8B	36.018	69.642	65.641	89.516	87.886	90.688
LATXA-70B	43.078	74.087	70.429	90.454	88.466	93.355
SFT ①	53.553	78.914	75.972	90.740	88.121	94.956
SFT ②	53.560	78.929	75.983	90.736	88.112	94.942
SFT ③	53.475	78.894	75.938	90.731	88.109	94.921
DPO$_{\text{SFT}}$ ①	53.635	78.959	76.019	90.751	88.133	94.891
DPO$_{\text{SFT}}$ ②	53.498	78.894	75.942	90.744	88.099	94.936
DPO$_{\text{SFT}}$ ③	53.713	78.997	76.060	90.765	88.137	94.911

(c) The 1,000 QE-extracted docs dataset.

Table 10: Full evaluation results against the testing datasets. For all metrics, higher is better. The best checkpoints are used for inference and evaluation across three independent runs to estimate confidence.

Table 11: Detailed analysis of snippets translated by the Llama-3.1-8B-Instruct model

English source	Basque reference	Basque translation	Errors		
Book about Bonifacio Etxegaray to be presented tomorrow in Zumaia.	Bonifazio Etxegarairen inguruko liburua aurkeztuko dute bihar Zumaian	Iparretako Bonifazio Etxegarai buruzko liburuaren aurkezpena ezkontoko Zumaian.	*“Iparretako”* - “from the North” added ADDITION	*“ezkontoko Zumaian”* - “in married Zumaia” / “Zumaia of the wedding” MISTRANSLATION MAJOR	The sentence is structured as a noun phrase “Presentation of the book about Bonifazio Etxegarai from the North in Zumaia.” SYNTAX
Lazaro Etxegaray will present the book “In search of Vasconia. The Life of Bonifacio Echegaray (1878-1956)” tomorrow at 19:00 in the Oxford hall.	’En busca de Vasconia. Vida de Bonifacio Echegaray (1878-1956)’ liburua aurkeztuko du bihar Lazaro Etxegaraik, 19:00etan Oxford aretoan.	Lazaro Etxegarairen ekarpena da “Vasconia bilatzen", "Bonifazio Etxegarai (1878-1956)ren bizitza” liburua gauzko 19:00etan, Oksfordgo aretokoan.	*“Lazaro Etxegarairen ekarpena da”* - “Lazaro Etxegarai’s contribution is” MISTRANSLATION MAJOR	*“gauzko”* doesn’t exist in Basque MISTRANSLATION MAJOR	
He first worked at the provincial library in Donostia and later became chronicler of the Basque Provinces.	Aldundiko liburutegian, Donostian, lanean aritu zen lehenbizi, eta Euskal Probintzietako kronista izatera pasa zen gero.	Lehenago Donostiako probintziako liburutegian lan egin zuen, eta geroa erdaldia egin zuen Euskal Probintzien erdaldia.	*“erdaldia”* does not exist in Basque MISTRANSLATION MAJOR	*“lehenago”* is wrong, should be *“lehenbizi”* GRAMMAR	
According to Lazaro, the documentation process was “difficult” because “information about the person was often found in archives outside the Basque Country”.	Lazaroren hitzetan, “zaila” izan da dokumentazio prozesua, “pertsonaiari buruzko informazioa askotan Euskal Herrian ez zeuden artxiboetan zegoelako”.	Lazaro-ren arabera, “informazioaren bilaketa zaila” izan zen “pertsona haren informazioa arkibategi gehienetan aurkitzen dela” gehienetan.	*“informazioaren bilaketa”* - “search for information” MISTRANSLATION MINOR	*“pertsona haren informazioa arkibategi gehienetan aurkitzen dela”* - “the person’s information is found in most archives” MISTRANSLATION MAJOR	
The Zaldibar Town Council and social organizations have demanded that the Basque Government take urgent measures regarding the landfill site.	Zaldibarko Udalak eta gizarte eragileek zabortegian premiazko neurriak hartzeko exijitu diote Jaurlaritzari	Gobernu Euskadiko Aldundia eta gizarte erakundeen artean, Zaldibarreko ontzi-utziaren inguruan hura begiratzeko mesedea eskatu dute.	*“Gobernu Euskadiko Aldundia”* - “Provincial Council of the Basque Government” MISTRANSLATION MAJOR	*“hura begiratzeko mesedea eskatu dute”* - “they have asked the favor to look at it” - missing “urgent measures” MISTRANSLATION MAJOR	*“ontzi-utziaren”* - “shipwreck” instead of “landfill site” MISTRANSLATION MAJOR
As they have explained, the site’s ground is becoming unstable, according to the latest geological report.	Azaldu dutenez, kezkatuta daude; izan ere, zabortegiko lurra egonkortasuna galtzen ari da, azken txosten geologikoaren arabera.	Elkarrekin esleitzen dutenez, ontzi-utziaren lurra desegoki bihurtzen ari da, zientzia geologiko baten arabera.	*“Elkarrekin esleitzen dutenez”* - “As they are allocated together” MISTRANSLATION MAJOR	*“ontzi-utziaren lurra”* - “land of the shipwreck” instead of “site’s ground” MISTRANSLATION MAJOR	
Therefore, they have demanded that the authorities provide explanations to the public.	Horregatik, egoeraren inguruko azalpenak herritarrei emateko eskatu dute.	`<missing>`	OMISSION		

Table 11: Detailed analysis of snippets translated by the Llama-3.1-8B-Instruct model (cont.)

English source	Basque reference	Basque translation	Errors		
Apart from being unpleasant, it's violent, being out in public, when scratch an itch, and your eye rotates, it goes white or pops out of its hole.	Ezatsegina izatetik aparte, biolentoa da, jendartean zaudela, kili-kilia duzulako hazka egin eta, begia biratuta, zuri jartzea edota zulotik kanporatzea.	Gaindikoa denbora eskainiko da, gauza bitxi, publikoan ariko baren bitxiagoa da, pixkaturik berrik, eta begia biratzen denean, zurrunbiloak edo begiaren urrezuluan ateratzen da.	"The excess will be given time, strange things, it's even stranger when I'm in public, a little bit new, and when the eye turns, the swirls or the eyeball comes out." MISTRANSLATION MAJOR		
My left eyelid is puffy, because they inserted silicone into it in a medical procedure to reduce the depth of the hollow and add volume to the area below it.	Ezker betazpia harrotuta daukat, saio mediko batean silikona sartu zidatelako botxoaren sakonera gutxitu eta betazpiak mamitzeko.	Ezkerreko begiaren azpian, kornealeko silicona sartu zuten, barru itxura askoaren urdibai alde batetik handituz, eta sakelako zuloaren oinaldean aldatuz, begi inguruak balantza gehitzen.	"Under the left eye, corneal silicone was inserted, enlarging the large, round-shaped pupil on one side and changing the base of the eye socket, adding volume to the eye area." MISTRANSLATION MAJOR		
In recent times, fascist violence has increased in the Spanish State, the Basque Country, and European countries.	Estatu Espainiarrean, Euskal Herrian eta Europako herrialdeetan, azken aldian, ugaritu egin da indarkeria faxista.	Izaera handitzeko denboraldian, faxismoa espainiar Estatuaren, Euskal Herrikoan eta herrialde europar batzuetan ere indarra hartzen ari da.	*"Izaera handitzeko denboraldian"* - "In the season for increasing character" instead of "In recent times" MISTRANSLATION MAJOR	*"faxismoa ... indarra hartzen ari da"* - "fascism is gaining strength" MISTRANSLATION MINOR	
Marx said that capitalism would die, but not on its own; it would need help to die, and that the bourgeoisie would not accept the death of capitalism, that it doesn't want capitalism to die, of course, as that would mean losing all their class privileges.	Marxek zioen kapitalismoa hil egingo dela, baina ez bere kabuz, hiltzen lagundu behar zaiola baina burgesiak ez duela onartuko kapitalismoaren heriotzarik, ez duela kapitalismoa hiltzerik nahi, noski, honekin batera bere klaseko pribilegio guztiak galduko lituzkeelako.	Marx esan zuen kapitalismoa hiltzea baino, bere aldean; laguntza izan behar zuen, eta burgesia kapitalismoaren heriotza onartzen ez zuen, klase privilegio guztiak galduko baitzituen.	*"Marx esan zuen kapitalismoa hiltzea baino, bere aldean"* - "Marx said rather than capitalism dying, on its side" GRAMMAR	missing "it would need help to die" OMISSION	*"burgesia ... onartzen ez zuen"* - "the bourgeoisie didn't accept" instead of "the bourgeoisie would not accept" GRAMMAR

Table 11: Detailed analysis of snippets translated by the Llama-3.1-8B-Instruct model (cont.)

English source	Basque reference	Basque translation	Errors		
It occurs when the capitalist regime changes and capitalists cannot rule as they wish. When there is a social crisis and the bourgeoisie can't protect its interests through "democratic" means, it makes use of fascist regimes.	Erregimen kapitalista aldatzean kapitalistek beti nahi duten moduan agindu ezin dezaketelako gertatzen da. Krisi sozial bat dagoenean, burgesiak bide "demokratikoekin" bere interesak babestu ezin dituenean erabili dituzte erregimen faxistak.	(hallucinate)	OMISSION		
To understand what fascism is, it is recommended to read Trotsky's "The Struggle Against Fascism".	Faxismoa zer den ulertu ahal izateko, gomendagarria da Trotskyren La Lucha contra el fascismo irakurtzea.	(hallucinate)	OMISSION		

Table 12: Detailed analysis of snippets translated by the Llama-3.1-70B-Instruct model

English source	**Basque reference**	**Basque translation**	**Errors**		
Book about Bonifacio Etxegaray to be presented tomorrow in Zumaia.	Bonifazio Etxegarairen inguruko liburua aurkeztuko dute bihar Zumaian	Liburua Bonifacio Etxegarayri buruz aurkeztuko da bihar Zumaian.			
Lazaro Etxegaray will present the book “In search of Vasconia. The Life of Bonifacio Echegaray (1878-1956)” tomorrow at 19:00 in the Oxford hall.	’En busca de Vasconia. Vida de Bonifacio Echegaray (1878-1956)’ liburua aurkeztuko du bihar Lazaro Etxegaraik, 19:00etan Oxford aretoan.	Lazaro Etxegarayk “Vasconia bila. Bonifacio Etxegarayren bizitza (1878-1956)” liburua aurkeztuko du bihar 19: 00etan Oxford aretoan.			
He first worked at the provincial library in Donostia and later became chronicler of the Basque Provinces.	Aldundiko liburutegian, Donostian, lanean aritu zen lehenbizi, eta Euskal Probintzietako kronista izatera pasa zen gero.	Lehendabizi, Donostiako liburutegi probintzialean lan egin zuen eta gero Euskal Probintzien kronikagilea izan zen.			
According to Lazaro, the documentation process was “difficult” because “information about the person was often found in archives outside the Basque Country”.	Lazaroren hitzetan, “zaila” izan da dokumentazio prozesua, “pertsonaiari buruzko informazioa askotan Euskal Herrian ez zeuden artxiboetan zegoelako”.	Lazaro Etxegarayren arabera, dokumentazio prozesua “zaila” izan zen “pertsonari buruzko informazioa askotan Euskal Herriko artxiboetan aurkitu baitzen”.	*"Euskal Herriko artxiboetan aurkitu baitzen"* - "because it was found in archives of the Basque Country" - should be "outside" MISTRANSLATION MAJOR		
The Zaldibar Town Council and social organizations have demanded that the Basque Government take urgent measures regarding the landfill site.	Zaldibarko Udalak eta gizarte eragileek zabortegian premiazko neurriak hartzeko exijitu diote Jaurlaritzari	Zaldibarko Udala eta erakunde sozialak eskatu dute Eusko Jaurlaritzak zabortegiari buruz neurri azkarrekin jardun behar duela.	*“eskatu dute . . . Jaurlaritzak . . . jardun behar duela”* - “they have asked that the Gov. must act” instead of “demanded” LEXICAL	*“neurri azkarrekin jardun”* - “act with quick measures” instead of “urgent” LEXICAL	
As they have explained, the site’s ground is becoming unstable, according to the latest geological report.	Azaldu dutenez, kezkatuta daude; izan ere, zabortegiko lurra egonkortasuna galtzen ari da, azken txosten geologikoaren arabera.	Azkeneko txosten geologikoaren arabera, zabortegiaren lurzorua ezegonkor bihurtzen ari dela adierazi dute.			
Therefore, they have demanded that the authorities provide explanations to the public.	Horregatik, egoeraren inguruko azalpenak herritarrei emateko eskatu dute.	Beraz, agintariek azalpenak eman ditzatela eskatu dute publikoari.	“Therefore, the authorities have asked the public to provide explanations.” MISTRANSLATION MAJOR		

Table 12: Detailed analysis of snippets translated by the Llama-3.1-70B-Instruct model (cont.)

English source	**Basque reference**	**Basque translation**	**Errors**		
Apart from being unpleasant, it's violent, being out in public, when scratch an itch, and your eye rotates, it goes white or pops out of its hole.	Ezatsegina izatetik aparte, biolentoa da, jendartean zaudela, kili-kilia duzulako hazka egin eta, begia biratuta, zuri jartzea edota zulotik kanporatzea.	Atseginarria izateaz gain, indarra da, publikoki, karraskatzeko gogoa sumatzen duzunean, eta begia biratzen zaizu, zuritzen zaizu edo zuloatik irteten zaizu.	"*Atseginarria*" - "pleasant" LEXICAL	"*indarra da*" - "it's a strength" LEXICAL	"*karraskatzeko gogoa sumatzen duzunean*" - "when you feel the urge to gnaw" LEXICAL
My left eyelid is puffy, because they inserted silicone into it in a medical procedure to reduce the depth of the hollow and add volume to the area below it.	Ezker betazpia harrotuta daukat, saio mediko batean silikona sartu zidatelako botxoaren sakonera gutxitu eta betazpiak mamitzeko.	Nire ezkerreko begiko begi-babeslea puztuta dago, zerikoa txertatu baitzuten zirujau batek begien itxura zuzentzeko eta begien azpiko gunea bolumenera eramateko.	"*begi-babeslea*" - "eye shield" LEXICAL "*zirujau batek*" - "a surgeon" added ADDITION	agreement is wrong GRAMMAR	"*begien itxura zuzentzeko eta begien azpiko gunea bolumenera eramateko*" - "to correct the eyes' appearance and bring the area below the eyes to volume" MISTRANSLATION MINOR
In recent times, fascist violence has increased in the Spanish State, the Basque Country, and European countries.	Estatu Espainiarrean, Euskal Herrian eta Europako herrialdeetan, azken aldian, ugaritu egin da indarkeria faxista.	Azkenaldian, faxismoaren indarkeria areagotu da Espainiako Estatuan, Euskal Herrian eta Europako herrialdeetan.			
Marx said that capitalism would die, but not on its own; it would need help to die, and that the bourgeoisie would not accept the death of capitalism, that it doesn't want capitalism to die, of course, as that would mean losing all their class privileges.	Marxek zioen kapitalismoa hil egingo dela, baina ez bere kabuz, hiltzen lagundu behar zaiola baina burgesiak ez duela onartuko kapitalismoaren heriotzarik, ez duela kapitalismoa hiltzerik nahi, noski, honekin batera bere klaseko pribilegio guztiak galduko lituzkeelako.	Marxek esan zuen kapitalismoa hiltzeko ez zuela bere kabuz hilko, laguntza beharko zuela hil ahal izateko eta burgesia ez zuela onartuko kapitalismoaren heriotza, ez zuela nahiela kapitalismoa hil, bere klase pribilegio guztiak galduko zituelako.			

Table 12: Detailed analysis of snippets translated by the Llama-3.1-70B-Instruct model (cont.)

English source	**Basque reference**	**Basque translation**	**Errors**		
It occurs when the capitalist regime changes and capitalists cannot rule as they wish. When there is a social crisis and the bourgeoisie can't protect its interests through "democratic" means, it makes use of fascist regimes.	Erregimen kapitalista aldatzean kapitalistek beti nahi duten moduan agindu ezin dezaketelako gertatzen da. Krisi sozial bat dagoenean, burgesiak bide "demokratikoekin" bere interesak babestu ezin dituenean erabili dituzte erregimen faxistak.	Gertatzen da kapitalismoaren erregimena aldatzen denean eta kapitalistek nahi bezala gobernatu ezin dutenean. Krise soziala dagoenean eta burgesiak ezin du "demokratiko" bidez bere interesei eusten, faxismoaren erregimenei erabiltzen ditu.	*"gobernatu"* - "govern" instead of *"agindu"* - "dictate" LEXICAL		
To understand what fascism is, it is recommended to read Trotsky's "The Struggle Against Fascism".	Faxismoa zer den ulertu ahal izateko, gomendagarria da Trotskyren La Lucha contra el fascismo irakurtzea.	Faxismoa zer den ulertzeko, Trotskiren "Faxismoaren aurkako borroka" irakurtzea gomendatzen da.			

Table 13: Detailed analysis of snippets translated by the SFT model

English source	Basque reference	Basque translation	Errors		
Book about Bonifacio Etxegaray to be presented tomorrow in Zumaia.	Bonifazio Etxegarairen inguruko liburua aurkeztuko dute bihar Zumaian	Bonifacio Etxegarayri buruzko liburua aurkeztuko dute bihar Zumaian			
Lazaro Etxegaray will present the book “In search of Vasconia. The Life of Bonifacio Echegaray (1878-1956)” tomorrow at 19:00 in the Oxford hall.	’En busca de Vasconia. Vida de Bonifacio Echegaray (1878-1956)’ liburua aurkeztuko du bihar Lazaro Etxegaraik, 19:00etan Oxford aretoan.	Lazaro Etxegarayk ‘Vasconiaren bila. Bonifacio Echegarayren bizitza (1878-1956)’ liburua aurkeztuko du bihar, 19:00etan, Oxford aretoan.			
He first worked at the provincial library in Donostia and later became chronicler of the Basque Provinces.	Aldundiko liburutegian, Donostian, lanean aritu zen lehenbizi, eta Euskal Probintzietako kronista izatera pasa zen gero.	Donostiako probintzia liburutegian aritu zen lehenik, eta Euskal Probintzietako kronista izan zen gero.			
According to Lazaro, the documentation process was “difficult” because “information about the person was often found in archives outside the Basque Country”.	Lazaroren hitzetan, “zaila” izan da dokumentazio prozesua, “pertsonaiari buruzko informazioa askotan Euskal Herrian ez zeuden artxiboetan zegoelako”.	Lazaro Etxegarayren esanetan, dokumentazioa lortzeko prozesua “zaila” izan da, “pertsonari buruzko informazioa Euskal Herritik kanpoko artxiboetan aurkitu behar izan dudalako askotan”.	nuance MISTRANSLATION MINOR		
The Zaldibar Town Council and social organizations have demanded that the Basque Government take urgent measures regarding the landfill site.	Zaldibarko Udalak eta gizarte eragileek zabortegian premiazko neurriak hartzeko exijitu diote Jaurlaritzari	Zaldibarko Udalak eta eragile sozialek zabortegiari buruzko premiazko neurriak hartzeko eskatu diote Eusko Jaurlaritzari	*“eskatu”* - “asked” instead of *“exijitu”* - “demanded” LEXICAL		
As they have explained, the site’s ground is becoming unstable, according to the latest geological report.	Azaldu dutenez, kezkatuta daude; izan ere, zabortegiko lurra egonkortasuna galtzen ari da, azken txosten geologikoaren arabera.	Azaldu dutenez, gunearen lurra ezegonkorra da, azken txosten geologikoaren arabera.	*“ezegonkorra da”* - “is unstable” instead of “is becoming unstable” GRAMMAR		
Therefore, they have demanded that the authorities provide explanations to the public.	Horregatik, egoeraren inguruko azalpenak herritarrei emateko eskatu dute.	Hori dela eta, agintariei azalpenak eman ditzatela eskatu diete.	*“herritarrei”* - “to the public” missing OMISSION		

Table 13: Detailed analysis of snippets translated by the SFT model (cont.)

English source	Basque reference	Basque translation	Errors		
Apart from being unpleasant, it's violent, being out in public, when scratch an itch, and your eye rotates, it goes white or pops out of its hole.	Ezatsegina izatetik aparte, biolentoa da, jendartean zaudela, kili-kilia duzulako hazka egin eta, begia biratuta, zuri jartzea edota zulotik kanporatzea.	Desatsegina izateaz aparte, bortitza da, jendaurrean egotea, when scratch an itch, and your eye rotates, it goes white or pops out of its hole.	latter half in English UNTRANSLATED		
My left eyelid is puffy, because they inserted silicone into it in a medical procedure to reduce the depth of the hollow and add volume to the area below it.	Ezker betazpia harrotuta daukat, saio mediko batean silikona sartu zidatelako botxoaren sakonera gutxitu eta betazpiak mamitzeko.	Nire ezkerreko betazala puztuta dago, because they inserted silicone into it in a medical procedure to reduce the depth of the hollow and add volume to the area below it.	latter half in English UNTRANSLATED		
In recent times, fascist violence has increased in the Spanish State, the Basque Country, and European countries.	Estatu Espainiarrean, Euskal Herrian eta Europako herrialdeetan, azken aldian, ugaritu egin da indarkeria faxista.	Azken aldian indarkeria faxista areagotu egin da Espainiar Estatuan, Euskal Herrian eta Europako herrialdeetan.			
Marx said that capitalism would die, but not on its own; it would need help to die, and that the bourgeoisie would not accept the death of capitalism, that it doesn't want capitalism to die, of course, as that would mean losing all their class privileges.	Marxek zioen kapitalismoa hil egingo dela, baina ez bere kabuz, hiltzen lagundu behar zaiola baina burgesiak ez duela onartuko kapitalismoaren heriotzarik, ez duela kapitalismoa hiltzerik nahi, noski, honekin batera bere klaseko pribilegio guztiak galduko lituzkeelako.	Marxek esan zuen kapitalismoa hil egingo zela, baina ez bere kabuz, hil ahal izateko laguntza beharko zuela, eta burgesiak ez duela onartuko kapitalismoaren heriotza, ez duela nahi kapitalismoa hiltzea, noski, horrek bere klase pribilegio guztiak galtzea ekarriko liokeelako.			
It occurs when the capitalist regime changes and capitalists cannot rule as they wish. When there is a social crisis and the bourgeoisie can't protect its interests through "democratic" means, it makes use of fascist regimes.	Erregimen kapitalista aldatzean kapitalistek beti nahi duten moduan agindu ezin dezaketelako gertatzen da. Krisi sozial bat dagoenean, burgesiak bide "demokratikoekin" bere interesak babestu ezin dituenean erabili dituzte erregimen faxistak.	Erregimen kapitalista aldatzen denean eta kapitalistek ezin dutenean nahi bezala gobernatu gertatzen da. Krisi soziala dagoenean eta burgesiak bere interesak bide "demokratikoen" bidez babestu ezin dituenean, erregimen faxistaz baliatzen da.	"*gobernatu*" - "govern" instead of "*agindu*" - "dictate" LEXICAL		

Table 13: Detailed analysis of snippets translated by the SFT model (cont.)

English source	Basque reference	Basque translation	Errors		
To understand what fascism is, it is recommended to read Trotsky's "The Struggle Against Fascism".	Faxismoa zer den ulertu ahal izateko, gomendagarria da Trotskyren La Lucha contra el fascismo irakurtzea.	Faxismoa zer den ulertzeko Trotskiren "Faxismoaren aurkako borroka" irakurtzea gomendatzen da.			

Table 14: Detailed analysis of snippets translated by the $\mathrm{DPO}_{\mathrm{SFT}}$ model

English source	Basque reference	Basque translation	Errors		
Book about Bonifacio Etxegaray to be presented tomorrow in Zumaia.	Bonifazio Etxegarairen inguruko liburua aurkeztuko dute bihar Zumaian	Bonifacio Etxegarayri buruzko liburua aurkeztuko dute bihar Zumaian			
Lazaro Etxegaray will present the book "In search of Vasconia. The Life of Bonifacio Echegaray (1878-1956)" tomorrow at 19:00 in the Oxford hall.	'En busca de Vasconia. Vida de Bonifacio Echegaray (1878-1956)' liburua aurkeztuko du bihar Lazaro Etxegaraik, 19:00etan Oxford aretoan.	Lazaro Etxegarayk 'Vasconiaren bila. Bonifacio Echegarayren bizitza (1878-1956)' liburua aurkeztuko du bihar, 19:00etan, Oxford aretoan.			
He first worked at the provincial library in Donostia and later became chronicler of the Basque Provinces.	Aldundiko liburutegian, Donostian, lanean aritu zen lehenbizi, eta Euskal Probintzietako kronista izatera pasa zen gero.	Donostiako probintzia liburutegian aritu zen lehenik, eta Euskal Probintzietako kronista izan zen gero.			
According to Lazaro, the documentation process was "difficult" because "information about the person was often found in archives outside the Basque Country".	Lazaroren hitzetan, "zaila" izan da dokumentazio prozesua, "pertsonaiari buruzko informazioa askotan Euskal Herrian ez zeuden artxiboetan zegoelako".	Lazaro Etxegarayren esanetan, dokumentazioa lortzeko prozesua "zaila" izan da, "pertsonari buruzko informazioa Euskal Herritik kanpoko artxiboetan aurkitu delako askotan".	nuance MISTRANSLATION MINOR		
The Zaldibar Town Council and social organizations have demanded that the Basque Government take urgent measures regarding the landfill site.	Zaldibarko Udalak eta gizarte eragileek zabortegian premiazko neurriak hartzeko exijitu diote Jaurlaritzari	Zaldibarko Udalak eta eragile sozialek zabortegiari buruzko premiazko neurriak hartzeko eskatu diote Eusko Jaurlaritzari	*"eskatu"* - "asked" instead of *"exijitu"* - "demanded" LEXICAL		
As they have explained, the site's ground is becoming unstable, according to the latest geological report.	Azaldu dutenez, kezkatuta daude; izan ere, zabortegiko lurra egonkortasuna galtzen ari da, azken txosten geologikoaren arabera.	Azaldu dutenez, gunearen lurra ezegonkorra da, azken txosten geologikoaren arabera.	*"ezegonkorra da"* - "is unstable" instead of "is becoming unstable" GRAMMAR		
Therefore, they have demanded that the authorities provide explanations to the public.	Horregatik, egoeraren inguruko azalpenak herritarrei emateko eskatu dute.	Hori dela eta, agintariei azalpenak eman ditzatela eskatu diete.	*"herritarrei"* - "to the public" missing OMISSION		

Table 14: Detailed analysis of snippets translated by the DPO$_{\text{SFT}}$ model (cont.)

English source	Basque reference	Basque translation	Errors		
Apart from being unpleasant, it's violent, being out in public, when scratch an itch, and your eye rotates, it goes white or pops out of its hole.	Ezatsegina izatetik aparte, biolentoa da, jendartean zaudela, kili-kilia duzulako hazka egin eta, begia biratuta, zuri jartzea edota zulotik kanporatzea.	Desatsegina izateaz aparte, bortitza da, jendaurrean egotea, when scratch an itch, and your eye rotates, it goes white or pops out of its hole.	latter half in English UNTRANSLATED		
My left eyelid is puffy, because they inserted silicone into it in a medical procedure to reduce the depth of the hollow and add volume to the area below it.	Ezker betazpia harrotuta daukat, saio mediko batean silikona sartu zidatelako botxoaren sakonera gutxitu eta betazpiak mamitzeko.	Nire ezkerreko betazala puztuta dago, because they inserted silicone into it in a medical procedure to reduce the depth of the hollow and add volume to the area below it.	latter half in English UNTRANSLATED		
In recent times, fascist violence has increased in the Spanish State, the Basque Country, and European countries.	Estatu Espainiarrean, Euskal Herrian eta Europako herrialdeetan, azken aldian, ugaritu egin da indarkeria faxista.	Azken aldian indarkeria faxista areagotu egin da Espainiar Estatuan, Euskal Herrian eta Europako herrialdeetan.			
Marx said that capitalism would die, but not on its own; it would need help to die, and that the bourgeoisie would not accept the death of capitalism, that it doesn't want capitalism to die, of course, as that would mean losing all their class privileges.	Marxek zioen kapitalismoa hil egingo dela, baina ez bere kabuz, hiltzen lagundu behar zaiola baina burgesiak ez duela onartuko kapitalismoaren heriotzarik, ez duela kapitalismoa hiltzerik nahi, noski, honekin batera bere klaseko pribilegio guztiak galduko lituzkeelako.	Marxek esan zuen kapitalismoa hil egingo zela, baina ez bere kabuz, hil ahal izateko laguntza beharko zuela, eta burgesiak ez duela onartuko kapitalismoaren heriotza, ez duela nahi kapitalismoa hiltzea, noski, horrek bere klase pribilegio guztiak galtzea suposatuko lukeelako.			
It occurs when the capitalist regime changes and capitalists cannot rule as they wish. When there is a social crisis and the bourgeoisie can't protect its interests through "democratic" means, it makes use of fascist regimes.	Erregimen kapitalista aldatzean kapitalistek beti nahi duten moduan agindu ezin dezaketelako gertatzen da. Krisi sozial bat dagoenean, burgesiak bide "demokratikoekin" bere interesak babestu ezin dituenean erabili dituzte erregimen faxistak.	Erregimen aldaketa hau gertatzen da diskurtso demokratikoak eta langile zuzendaritzak ez direnean kapitalismoaren existentzia zalantzan jartzen duen langile mugimendu gero eta indartsuagoa gelditzeko gai.	"*gobernatu*" - "govern" instead of "*agindu*" - "dictate" LEXICAL		

Table 14: Detailed analysis of snippets translated by the DPO_{SFT} model (cont.)

English source	Basque reference	Basque translation	Errors		
To understand what fascism is, it is recommended to read Trotsky's "The Struggle Against Fascism".	Faxismoa zer den ulertu ahal izateko, gomendagarria da Trotskyren La Lucha contra el fascismo irakurtzea.	Faxismoa zer den ulertzeko Trotskiren "La lucha contra el fascismo" irakurtzea gomendatzen da.			