CUNI at WMT24 General Translation Task: LLMs, (Q)LoRA, CPO and Model Merging

Miroslav Hrabal, Josef Jon, Martin Popel, Nam H. Luu, Danil Semin, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics
{hrabal,jon,popel,bojar}@ufal.mff.cuni.cz,
namhoang.luu700@student.cuni.cz, dsemin2305@gmail.com

Abstract

This paper presents the contributions of Charles University teams to the WMT24 General Translation task (English to Czech, German and Russian, and Czech to Ukrainian) and the WMT24 Translation into Low-Resource Languages of Spain task. Our most elaborate submission, CUNI-MH for en2cs, is the result of fine-tuning Mistral 7B v0.1 for translation using a three-stage process: Supervised fine-tuning using QLoRA, Contrastive Preference Optimization, and merging of model checkpoints. We also describe the CUNI-GA, CUNI-Transformer and CUNI-DocTransformer submissions, which are based on our systems from the previous year.

Our en2ru system CUNI-DS uses a similar first stage as CUNI-MH (QLoRA for en2cs) and follows with transfer learning for en2ru.

For en2de (CUNI-NL), we experimented with an LLM-based speech translation system, to translate without the speech input.

For the Translation into Low-Resource Languages of Spain task, we performed QLoRA fine-tuning of a large LLM on a small amount of synthetic (backtranslated) data.

1 Introduction

This paper describes the CUNI submissions to the WMT24 General Translation task (from English to Czech, German and Russian, and from Czech to Ukrainian) and the Translation into Low-Resource Languages of Spain task.

Our underlying goal for this year was to test the applicability of primarily small open-source LLMs to the languages of interest, and we also provide our English-to-Czech systems from the previous years for comparison.

The setups for the various target languages differ considerably in the methods used. Table 1 provides an overview of the individual system highlights. In Section 2, we detail the basic building steps and methods across our systems (not all setups use all of them). Section 3 describes the training and development data used across the target languages. In Section 4, we evaluate the systems and compare their results with various available baselines and benchmarks. Section 5 summarizes our future plans, and we conclude in Section 6.

2 Methods

For the CUNI-MH submission, we fine-tuned Mistral 7B v0.1 (Jiang et al., 2023) using three stages:

1. Supervised fine-tuning on CzEng 2.0 training dataset (Kocmi et al., 2020)[1], see Section 2.3.
2. Contrastive Preference Optimization (Xu et al., 2024b), see Section 2.4.
3. Averaging model checkpoints (Utans, 1996; Wortsman et al., 2022; Gueta et al., 2023), see Section 2.5.

CUNI-Transformer and CUNI-DocTransformer are the same systems as submitted last year (Jon et al., 2023), relying on standard NMT training with Block backtranslation (Section 2.1) and optionally document-level training (Section 2.2).

For CUNI-GA, in English-to-Czech, we used outputs from CUNI-Transformer and a genetic algorithm to combine and modify them, again in the same way as previous year (Section 2.8; Jon et al., 2023; Jon and Bojar, 2023). For coincidentally identically called CUNI-GA submission in Translation into Low-Resource Languages of Spain task, we fine-tune larger LLMs (Command-R and Aya-23), without applying the genetic algorithm.

For the CUNI-NL system, we fine-tuned Llama 2 7B (Touvron et al., 2023) for the speech translation task, while also adapting it for text-only translation at the same time; see Section 2.6.

Finally CUNI-DS starts as step 1 of CUNI-MH but continues with transfer learning to target Russian instead of Czech, see Section 2.7.

[1] http://ufal.mff.cuni.cz/czeng/

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 232–246
November 15–16, 2024. ©2024 Association for Computational Linguistics

Task	CUNI-* Model	Initial LLM	SFT Data	SFT Highlights (§2.3)	Final Stages
cs2uk	Transformer	-	Opus, CzEng	BlockBT §2.1	-
en2cs	DocTransformer	-	CzEng 2.0	BlockBT §2.1, doc-level §2.2	-
en2cs	GA	-	-	-	GA §2.8
en2cs	MH	Mistral 7B v0.1	CzEng 2.0	QLoRA, Packing, AdamW	CPO §2.4; Checkpoint Merging §2.5
spa	GA	Command-R, Aya	PILAR BT	QLoRA	-
en2de	NL	HuBERT, Llama-2-7b	MuST-C	Text-only use of a speech translation system §2.6	
en2ru	DS	Mistral 7B v0.1	CzEng, Yandex, News Commentary	Transfer from en2cs §2.7	-

Table 1: Overview of CUNI systems in WMT24 General Translation task and Translation into Low-Resource Languages of Spain task (spa). Systems in the upper part of the table are our last year's baselines. §· refer to the methods in Section 2.

2.1 BlockBT

For training CUNI-Transformer and CUNI-DocTransformer, we used iterated Block backtranslation (BlockBT) (Popel, 2018; Popel et al., 2020; Gebauer et al., 2021; Jon et al., 2022) in a standard Transformer (Vaswani et al., 2017) NMT training from scratch. The BlockBT method organizes the training data, so that the model can optimize the balance between authentic English-to-Czech parallel texts (exhibiting more translationese artifacts) and synthetic data created by back-translating Czech-only texts) by averaging eight checkpoints reflecting more of the former or the latter domain. The use of eight checkpoints for averaging is derived from the original paper (Popel, 2018) and a study on hyperparametrs for training Transformers (Popel and Bojar, 2018).

2.2 Document-level training

The approach for training CUNI-DocTransformer is described in Popel et al. (2019). Starting with the initial sentence-level model (CUNI-Transformer), we continued training on sequences of consecutive sentences coming from a coherent text with at most 3000 characters, where both sides (en and cs) have the same number of sentences. The sentences are separated by a special token in each of the languages.

2.3 Supervised fine-tuning (SFT)

For the CUNI-MH submission, we used 4-bit QLoRA (Dettmers et al., 2023) with a large LoRA rank of $r = 512$. We used a batch size of 32, a learning rate of $2e-5$, 20 warm-up steps, 8-bit AdamW (Loshchilov and Hutter, 2019) optimizer and weight decay of 0.01. We also used a scheduler with linear learning rate decay. Starting from the freely available Mistral 7B v0.1 model, we trained in a language modeling fashion on individual sentences, calculating the loss on each token. To reduce the number of padding tokens, we also used packing: examples are concatenated with the EOS token as a separator to achieve a total sequence length of 1000. In Appendix A, we present our translation prompt template and example of its processed form with packing as used during training.

We trained for a single epoch on the authentic part of CzEng 2.0. In Figure 1, we show how the performance of the model develops during the first stage, starting from 100 steps. A notable observation is that the COMET22 and COMETKIWI22 scores seem to plateau relatively early, despite the evaluation loss steadily decreasing, while BLEU seems to be steadily increasing. This appears to be consistent with the results presented by Xu et al. (2024a), although we suspect it could also be the result of insufficient regularization.

For training, we used the HuggingFace Transformers and TRL libraries by Wolf et al. (2020) and von Werra et al. (2020). We also used the Unsloth library,[2] which provides speed and VRAM optimizations to Transformers and TRL libraries.

Another of our submissions that made use of a pre-trained LLM and SFT was CUNI-GA in the Translation into Low-Resource Languages of Spain task. We used 4-bit QLoRA with the rank of $r = 16$ and the learning rate of $4e-4$ for fine-tuning the pretrained *Command-R* model, and $1e-3$ for fine-tuning the *Aya* model, with an effective batch size

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

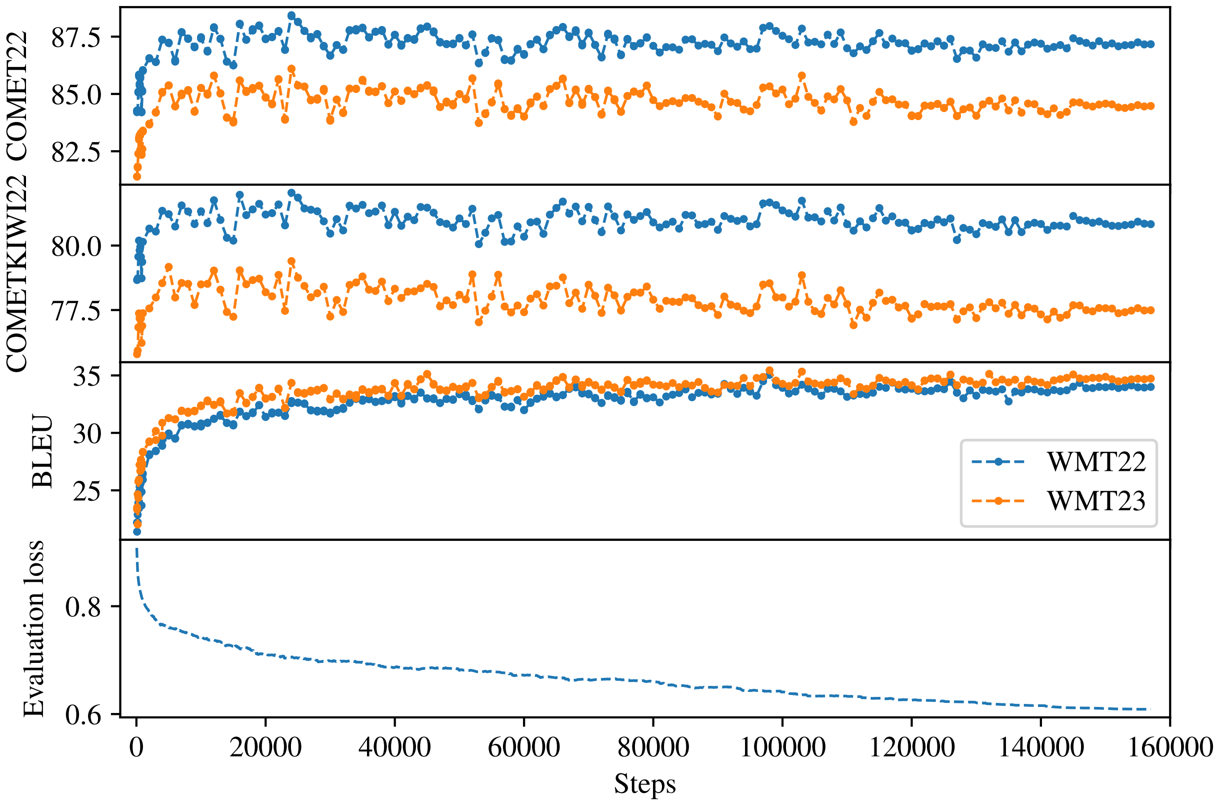

Figure 1: CUNI-MH Stage 1 – metrics during training.

of 32 and an AdamW optimizer with the weight decay value of 0.001.

2.4 Contrastive Preference Optimization (CPO)

CPO is a fine-tuning method introduced by Xu et al. (2024b) as an approximation of Direct Preference Optimization (Rafailov et al., 2024).

The goal of CPO is to fine-tune the model to directly optimize for preferences between translation candidates, rather than just optimizing the likelihood of the reference translations.

From a high-level point of view, the main difference between using SFT and CPO for translation is that for a given source text, we need two translations: *preferred* and *dis-preferred*. This means that the training dataset consists of triplets, rather than pairs as is typical for supervised training of NMT. For a more detailed description of the dataset we used and how it was created, see Section 3.2.

To apply CPO during the second stage of CUNI-MH training, we started two separate training runs from models we created during the first stage. One of the runs starts from model 3 and the other from model 4 in Table 2.

We selected these models because they had the best COMET22 and COMETKIWI22 scores among the models we had available at the time, when evaluated on the sentence-level WMT22 validation set.

Because we wanted to use a smaller LoRA rank size comparable to those used in the original paper (Xu et al., 2024b), we merged LoRA adapters with the quantized model into a 16-bit model and added new, smaller adapters.

We trained for two epochs with the following parameters: LoRA rank $r = 32$, LoRA $\alpha = 64$, CPO $\beta = 0.1$. We trained two separate runs, starting from the checkpoints mentioned earlier. Similarly to the SFT stage, we used 8-bit AdamW, this time without learning rate decay. Our GPU memory capacity was limiting us to the batch size of 4, so to compensate, we used 64 gradient accumulation steps to simulate a larger effective batch size of 256.

Stage	ID	Model	Checkpoint	COMET22	COMETKIWI22	BLEU
	0	Mistral 7B v0.1 5-shot		67.16	59.79	17.35
1	1	SFT from 0	16000	85.59	79.04	33.46
1	2		24000	86.10	79.40	34.35
1	3		103000	85.80	78.85	35.32
1	4	SLERP merge of 1 and 2		86.16	79.44	35.15
2	5	CPO from 4	150	89.76	82.71	32.56
2	6	CPO from 3	100	89.93	83.04	34.43
2	7	CPO from 0	400	83.21	76.54	18.33
3	**8**	**Linear merge of 5 and 6**		**90.21**	**83.16**	**36.52**

Table 2: CUNI-MH's training stages, models and their sentence-level scores on WMT23 (test set). The final CUNI-MH submission 8 is in bold.

Checkpoints were saved every 50 steps[3] and evaluated on the validation test set using COMETKIWI22. The performance peaked around checkpoint 150 for the first run, leading us to conclude that further training beyond 2 epochs was unnecessary. However, we acknowledge that the training parameters may not be optimal and could potentially be tweaked further for better results.

2.5 Checkpoint merging

To further improve the performance of the CUNI-MH model, we experimented with two methods for merging model weights: linear interpolation (Utans, 1996) and spherical linear interpolation (SLERP, Shoemake, 1985) in different training stages.

In particular, after the SFT stage, we merged two promising checkpoints from the same training run using SLERP, which led to a small improvement in all metrics, as can be seen by looking at model 4 in Table 2.

After the CPO stage, we once again experimented with model merging, this time we merged the best performing checkpoints from two different CPO training runs. This led to a further modest improvement in all COMET22, COMETKIWI22 and BLEU metrics, as shown by model 8 in Table 2.

For model merging using both SLERP and linear interpolation, we used the `mergekit` library by Goddard et al. (2024).

2.6 SFT from Speech Translation System (SFTSpeech)

The CUNI-NL system was adapted from a speech translation system, which features a frozen HuBERT component (Hsu et al., 2021) and the Llama 2 7B (Touvron et al., 2023) LLM.

The original speech translation system applied the CTC collapsing strategy to extract the speech hidden features; these features would subsequently be given as the prompt to a LLM to generate the ASR transcription and its corresponding translation simultaneously.

For the purposes of the General Translation Task, we avoid any audio features during inference and directly prompt the LLM with the source language text. We expect the LLM to translate using that only information. The motivation for this experiment was to check if a LLM-based speech translation system remains versatile enough to support text-only translation.

The original speech translation system was a fine-tuned LLM using 4-bit QLoRA (Dettmers et al., 2023) adapters, with the rank of $r = 8$ and alpha of $\alpha = 8$. Other training hyperparameters included the batch size of 1, the learning rate of $1e-4$ with 10 warmup steps, and an AdamW optimizer (Loshchilov and Hutter, 2019) with a cosine scheduler (Loshchilov and Hutter, 2017).

2.7 SFT for Transfer Learning

We used transfer learning across languages in the CUNI-DS system for English-to-Russian, transferring from English-to-Czech system.

2.7.1 Phase 1: en2cs Training

In the first phrase, we proceeded very similarly as described in Section 2.3. We started with the 4-bit quantized Mistral 7B v0.1 model (Jiang et al., 2023) and trained it using QLoRA (Dettmers et al., 2023) with a rank of 64 and an alpha of 128. The training followed Alpaca-like (Taori et al., 2023)

[3]Resulting in total of 7 checkpoints for each of the two runs.

instructions, with 20 warmup steps, a learning rate of $2e-5$, weight decay of $1e-2$, and a cumulative batch size of 32.

The model was trained on CzEng 2.0 for 24 hours, with segments packed into chunks of 2048 tokens. The final checkpoint was selected for the next phase.

2.7.2 Phase 2: en2ru Fine-Tuning

The model was then fine-tuned for en2ru translation using the Yandex Corpus for sentence-level data and the News Commentary v18.1 dataset for paragraph-level data. The datasets were shuffled and concatenated, and fine-tuning was conducted under the same conditions as the first stage, lasting 24 hours.

2.8 Genetic algorithm

For the CUNI-GA submission in English-to-Czech, we used a genetic algorithm to combine and modify n-best lists (Jon and Bojar, 2023) produced by CUNI-Transformer (at the sentence level), in the same manner as in Jon et al. (2023). We combined 5 metrics for the fitness function by a weighted average: BLEU (Papineni et al., 2002), chrF (Popović, 2015), wmt22-comet-da (Rei et al., 2022a), wmt22-cometkiwi-da (Rei et al., 2022b) and wmt23-cometkiwi-da-xl (Rei et al., 2023). The reference-based metrics use MBR decoding (Freitag et al., 2022) in place of the unknown reference.

3 Data

This section details the dataset used across the various training steps and language pairs.

3.1 SFT dataset

3.1.1 English-Czech

For the first stage of the CUNI-MH training, we used the authentic part CzEng 2.0. We did not use any preprocessing, except for applying the prompt template and packing described in Appendix A.

3.1.2 English-German

The CUNI-NL system was trained using the MuST-C dataset (Cattoni et al., 2021), a large multilingual corpus built from English TED Talks, containing the audio data, the English transcription of such audio, with its translation in multiple languages. Specifically, we used the en2de subset, consisting of approximately 400 hours of speech data.

During training, we randomly took 25% of the dataset, in which the input was the source transcript itself, instead of the audio features, so that the system could know how to translate from text-only data.

We trained the system for two epochs, both checkpoints of which were then used for evaluating against the WMT23 test set.

3.1.3 English-Russian

The initial phase of CUNI-DS system training (en2cs) utilized the first million segments from the CzEng 2.0 (Kocmi et al., 2020) dataset. In the second phase (en2ru), a combination of the Yandex Corpus[4] and the News Commentary v18.1[5] dataset was used, with the latter segmented into chunks of 10 sentences each.

3.1.4 Translation into Low-Resource Languages of Spain

For the Translation into Low-Resource Languages of Spain task, we backtranslated the literary part (`literary.txt`) of the PILAR dataset (Galiano-Jiménez et al., 2024) into Spanish using Apertium (Forcada and Tyers, 2016), resulting in 230k, 25k and 24k sentence pairs for Aranese, Aragonese and Asturian, respectively. For Aranese, we also backtranslated the Aranese side of the parallel part of the corpus, while keeping the paragraphs whole up to the length of 30 sentences, resulting in 726k sentences in 4329 documents. To make use of the paragraph-level context, we employed a context-aware prompt shown in Appendix B.

3.2 CPO dataset

To create a dataset for CPO (Section 2.4), we need triplets: source segment, preferred output and dispreferred output. We construct these triplets at the *paragraph level* (i.e. several sentences concatenated into a single segment) but sentence-level processing, inspired by the approach of (Xu et al., 2024b), is used in the preparation as described below.

Given a source segment, we select both preferred and dis-preferred translation from three candidates: our stage 1 output, our last year's constrained system and human reference. Our approach ensures that we still satisfy the requirements for a constrained submission.

Our CPO source segments (and their corresponding manual reference translations) are ran-

[4] https://translate.yandex.ru/corpus?lang=en

[5] https://data.statmt.org/news-commentary/v18.1/

Source text	Preferred translation	Dis-preferred translation
E6 goes further north along the west coast and through Norway to the Norwegian town Kirkenes at Barents Sea.	E6 pokračuje dále na sever podél západního pobřeží a přes Norsko do norského města Kirkenes u Barentsova moře.	E6 pokračuje dále na sever podél západního pobřeží a přes Norsko do norského města Kirkenes v Barentsově moři.
He became seriously ill in October 1914 and retired.	V říjnu 1914 vážně onemocněl a odešel do důchodu.	V říjnu 1914 ∅ onemocněl a odešel do důchodu.
This was published in June 1925, in a special issue of Poetry magazine.	Tato báseň byla publikována v červnu 1925 ve speciálním vydání časopisu Poetry.	Ta vyšla v červnu 1925 ve zvláštním čísle časopisu Poezie.
This convention has been ratified and acceded to by Ghana.	Tuto úmluvu ratifikovala a přistoupila k ní Ghana.	Tato úmluva byla ratifikována a přistoupena k ní Ghana.

Table 3: Short examples from the CPO dataset. Errors (underlined) are, resp.: Kirkenes located *in* Barents Sea; missed the adverb *seriously*; and grammatically inacceptable form of passivization mentioning the subject Ghana. The third example's dis-preferred translation does not mention the detail that we are referring to a poem ("báseň"), although this fact is not explicit in the source either; other lexical variations are minor.

domly sampled documents from CzEng 2.0, a total of 47257 documents containing 200k sentences. We then used the best checkpoint from stage 1 (see model 4 in Table 2) together with our constrained model from the previous year, CUNI-DocTransformer, to generate translations for the samples.[6]

Because we want to consider the manual translation as one of the candidates for the (dis-)preferred translation, we cannot use it as the reference to select the better candidate. Therefore, we use the reference-free wmt20-comet-qe-da[7] model to rank the translations, selecting the one with the highest score as the preferred one and the one with the lowest score as the dis-preferred one.

Note that wmt20-comet-qe-da scores individual sentences, not complete paragraphs, so we do this for each sentence in the sampled dataset, while giving all preceding sentences in the corresponding document (as translated by the given system) as a context (DocCOMET, Vernikos et al., 2022).

Since this DocCOMET approach is currently not supported by the COMET project[8] for newer model architectures, such as those used by COMETKIWI22 and XCOMET, we have not tried to build the data set using these newer models.

To arrive back at paragraph-level segments for CPO, we concatenate all the sentences in each original document. The result is a dataset consisting of 47k paragraph-level triplets for CPO. Each triplet consists of the paragraph in source language and two translations: preferred[9] and dis-preferred.[10] Due to the sentence-level selection, both preferred and dis-preferred translations may actually mix sentences from each of the three seed translations: human, our CUNI-DocTransformer and CUNI-MH Stage 1. We leave the analysis of document-level errors that arise in this process for future.

In Figure 2, we show which sentences were selected as preferred and dis-preferred. Note that this comparison is done on sentence-level, because the resulting paragraph-level examples can be composed of sentences from different sources. Interestingly, reference sentences were scored lowest by wmt20-comet-qe-da most frequently. We also show a few short examples from our dataset in Table 3. During training, the source sentences are formatted with the prompt template shown in Appendix A, similarly to how they are handled in the SFT stage Section 2.3.

We are aware that there are several potential issues with our method of preparing the dataset. First, there is a reason to be concerned about potential overfitting to a given metric (wmt20-comet-qe-da in our case) used to select the sentences. Second, our stage 1 CUNI-MH model did the translation in sentence-level fashion, potentially disregarding the relevant context. Third, we select sentences for preferred vs. dis-preferred class considering their preceding source-side context and their preceding target-side context as translated by the candidate system, not as selected so far within the document. This leaves document-level properties both in the positive and negative cases unhandled. Ideally, the preferred paragraph would avoid also any contextual errors, and for the dis-preferred paragraph, we

[6]For clarity, we note that we create only one CPO dataset, using translations by 4, and we apply the CPO method using this dataset three times, starting from three different models, see Table 2.

[7]https://huggingface.co/Unbabel/wmt20-comet-qe-da

[8]https://github.com/Unbabel/COMET

[9]Sometimes also called chosen or positive example.

[10]Sometimes also called rejected or negative example.

Model	COMET22	COMETKIWI22	BLEU
CUNI-Transformer	87.19	80.45	41.44
CUNI-DocTransformer	88.29	81.32	42.47
CUNI-GA	90.78	84.43	43.27
GPT4-5shot	89.36	82.82	37.76
CUNI-MH	90.21	83.16	36.52

Table 4: CUNI-MH's sentence-level scores on the en2cs WMT23 test set. Other systems' scores are taken from WMT23's automatic evaluation results.

Model	COMET22	COMETKIWI22	BLEU
CUNI-Transformer	81.13	68.24	42.27
CUNI-DocTransformer	83.52	70.69	43.29
CUNI-GA	86.15	73.56	43.83
GPT4-5shot	85.45	72.57	38.45
CUNI-MH $k = 1$	87.35	73.30	37.47
CUNI-MH $k = 8$	**87.73**	**74.82**	**35.42**

Table 5: CUNI-MH's document-level scores on the en2cs WMT23-para test set. k denotes how many sentences at most are translated together in one chunk. The CUNI-MH final submission is in bold.

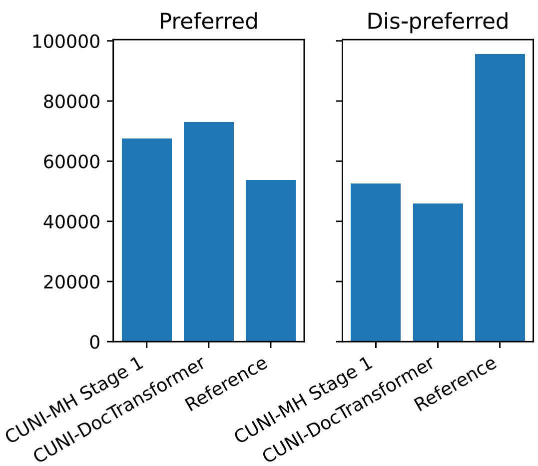

Figure 2: CPO dataset - sources of preferred and dis-preferred translations.

could construct worse translations in two ways: (1) using worse individual segments, as we do, and (2) combining better or worse individual segments in a way that purposefully damages paragraph context. Fourth, because we sampled uniformly from the CzEng 2.0 documents, our final dataset actually has a large number of documents, namely 24744 out of 41835, that only consist of a single sentence. We opted for a trivial sampling because we were concerned that naive solutions aiming at having more longer documents could potentially have a negative impact on the diversity of the dataset, however this is something we would like to address in the future.

All in all, we believe that there is potential to make subsequent iterations of the dataset higher quality by alleviating some of these concerns.

3.3 Validation and test datasets

During training of CUNI-MH, we used the WMT22 test set as the validation data set and the WMT23 test set as the test data set. In particular, we used WMT22 when selecting the best checkpoints and hyperparameters and only used WMT23 to estimate the final performance compared to baselines.

To prepare for paragraph-level evaluation, we also concatenated all the sentences in each document to a long paragraph, creating what we call WMT22-para and WMT23-para data sets. For CUNI-GA in English-to-Czech, we did not use validation sets, we did not compare the possible configurations on validation set, we chose the parameters based on our experience. For CUNI-GA in Translation into Low-Resource Languages of Spain, we use FLORES+ validation set (NLLB Team et al., 2022).

4 Evaluation

4.1 English-Czech

We show the sentence-level metrics on the WMT23 test set for the CUNI-MH system in Table 4 and the document-level metrics on the WMT23 test set in Table 5. We used greedy decoding for this system.

Since our preliminary experiments on WMT22-

Submission	W_{BLEU}	W_{CHRF}	W_{CMT22}	W_{QE22}	$W_{QE23-XL}$	CHRF	BLEU	QE22	QE23-XL	MetricX
CUNI-Transformer	-	-	-	-	-	57.3	29.3	X	0.614	4.3
CUNI-GA	0.1	0.1	0.4	0.4	0	56.4	29.5	0.819	0.658	-
CUNI-GA	0	0	0.5	0.5	0	55.5	26.5	0.827	0.650	-
CUNI-GA	**0**	**0**	**0.5**	**0**	**0.5**	**54.8**	**25.6**	**0.797**	**0.726**	**3.7**

Table 6: Paragraph-level scores on WMT24 test set for the CUNI-GA submission, primary submission in bold. CUNI-Transfomer was used to produce the n-best lists which are combined and modified for the CUNI-GA submission.

Model	COMET22	COMETKIWI22	BLEU
Baseline	24.04	28.55	0.20
CUNI-NL (epoch=1)	81.07	77.23	29.61
CUNI-NL (epoch=2)	80.90	77.51	30.75

Table 7: CUNI-NL's sentence-level scores on the en2de WMT23 test set.

para showed that our model did not handle longer paragraphs or documents well, we used sentence-splitter from Moses[11] to split segments into sentences. We then concatenate these sentences into chunks of up k, which we translate together as a whole. We then concatenate all the chunks to the original segments.

By testing our model on the WMT22-para validation dataset, we chose to use $k = 8$ for our final submission to optimize for the highest COMET22 and COMETKIWI22 scores. This can also be seen in Table 5, where the model with $k = 8$ has better COMET22 and COMETKIWI22 scores than the one with $k = 1$, at the cost of worse BLEU score.

The submitted CUNI-MH system also seems to perform well according to the preliminary automatic rankings, where it surpasses most of our systems from previous years and closely matching the performance of another of our systems, CUNI-GA. These results are shown in Table 8.

However, since both systems use COMET or COMETKIWI metrics during either training or inference, raising potential concerns about overfitting, we are also awaiting the results of human evaluation (Kocmi et al., 2024).

We also tried to use CPO with our new dataset to train the base Mistral model directly, skipping the supervised fine-tuning stage. The results are shown in Table 2, see 7, which is the best performing checkpoint of the training run, according to its COMETKIWI22 score on the validation dataset. It can be seen that the performance of this model is significantly worse in all metrics, so the SFT stage seems necessary in our setting.

We have also submitted CUNI-Transformer and CUNI-DocTransformer systems from previous year to provide reasonable constrained baselines for our newer models.

The CUNI-GA in this task submission combines hypotheses from CUNI-Transformer n-best lists created with beam sizes 4, 10 and 25 for each sentence. The resulting 39 translation candidates were processed by the genetic algorithm. The fitness (objective) function was a weighted combination of 5 metrics: BLEU, chrF, wmt22-comet-da (CMT22 in Table 6), wmt22-cometkiwi-da (QE22) and wmt23-cometkiwi-da-xl (QE23-XL). The weights and the obtained scores (chrF, BLEU, QE22, QE23-XL and MetricX (Juraska et al., 2023)) on the WMT24 test set are shown in Table 6. We did not use a development set due to high computational requirements of this approach, the weights are chosen based on our previous experience. An expected conclusion is that our approach allows us to easily optimize for the fitness metrics, which can be seen by comparing the QE23-XL scores of baseline translations (first row) and the score of the translations directly optimized for this metric (last row).

4.2 Czech-Ukrainian

We will add results for the Czech-Ukrainian submission in the camera-ready version.

4.3 English-German

For the CUNI-NL submission, we performed inference using the beam search algorithm, with the beam size of 2 for both checkpoints. We evaluated the performance of the two checkpoints of this system (as trained for speech translation), after epoch

[11] Wrapped by https://pypi.org/project/mosestokenizer/

English-Czech

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation?
Unbabel-Tower70B	1.0	1.8	0.732	✓
Claude-3.5 §	2.1	2.4	0.693	✓
CUNI-MH	2.1	2.3	0.690	✓
CUNI-GA	2.3	3.7	0.726	✓
Gemini-1.5-Pro	2.6	2.8	0.678	✓
GPT-4 §	2.6	2.9	0.682	✓
IOL-Research	2.8	3.0	0.676	✓
ONLINE-W	2.8	2.8	0.669	✓
CommandR-plus §	2.9	2.9	0.669	✓
SCIR-MT	3.2	3.3	0.664	✓
TranssionMT	3.5	3.5	0.655	
ONLINE-A	3.6	3.4	0.648	
Mistral-Large §	3.7	3.6	0.647	
IKUN	3.9	3.7	0.638	✓
ONLINE-B	4.0	3.9	0.640	
Llama3-70B §	4.1	4.0	0.640	✓
Aya23	4.3	4.0	0.630	✓
CUNI-DocTransformer	4.4	4.0	0.621	✓
IKUN-C	4.7	4.3	0.618	✓
CUNI-Transformer †	4.7	4.3	0.614	
ONLINE-G	5.7	5.2	0.592	
NVIDIA-NeMo †	7.6	6.5	0.536	
Phi-3-Medium §	15.0	11.4	0.305	
TSU-HITs	19.5	16.6	0.235	
CycleL2	24.2	19.5	0.077	
CycleL	27.0	22.5	0.031	

Table 8: Preliminary WMT24 General MT automatic ranking for English-Czech. **Closed systems** are highlighted with a dark gray background, **open systems** with a light gray background, and **constrained systems** are shown on a white background.

1 and after epoch 2 of en2de MuST-C corpus, with the latter performing better, so we chose it for the final evaluation against the test set this year. The results of the evaluation on the WMT23 test set are shown in Table 7.

4.4 English-Russian

For the CUNI-DS submission, we ran the evaluation on the paragraph level, i.e. the model needed to output the translation of the whole input at once. We used greedy decoding due to frequent emission of repeated tokens (sometimes called "spasm" by NMT practitioners) we observed with beam search. The outcomes of the CUNI-DS system's two-stage training are presented in Tables 9 and 10.

4.5 Translation into Low-Resource Languages of Spain

We compare Apertium and two open-source LLMs – Aya-23-8B and Command-R (35B version, quantized to 4 bits) – in translation from Spanish into the other languages of the task. We show the scores in Table 11. We fine-tuned both LLMs as a single joint model for all the languages on the back-translated literary data described in Section 3. We present BLEU, chrF and COMET-22 scores of the best-performing checkpoints after fine-tuning in Table 12. We submitted the translations produced using the Aya-23 model fine-tuned for 5000 steps. While the results are at best comparable to Apertium scores, we note that we only did a very light-weight fine-tuning on synthetic (backtranslated) data, which shows the potential of LLMs for translation into previously unsupported low-resource languages related to a language present in the training data. For instance, we obtained improvement from 46.7 to 70.2 ChrF (12.4 to 39.0 BLEU) in Aragonese by fine-tuning on 24k backtranslated sentence pairs from a different (literary) domain.

5 Future work

We have several ideas to improve the performance of the future iterations of our CUNI-MH model:

- Longer sequences: During our SFT stage, we trained on short sequences, mostly single sentences. In the future, we would like to experiment with training on larger sequences, so that the model is able to handle longer inputs in end-to-end fashion.
- Better CPO dataset: Our current dataset for CPO (Section 3.2) was created without including any filtering steps. The Stage 1 model we used to create one kind of translation candidates also translated in sentence-level fashion only. We think there is potential to create a higher quality dataset by using our final model, ensuring all translations are done with paragraph or document level context and possibly investigating means of filtering out lower quality examples.
- Better QLoRA initialization: During our SFT stage, we used the default initialization from the original LoRA paper (Hu et al., 2021). There are other initialization methods specifically for the combination of LoRA adapters and quantization, such as LoftQ (Li et al., 2023) which seems to consistently perform better for QLoRA. In the future, we would like to evaluate using this initialization method.
- Monolingual pretraining stage: Xu et al. (2024a) have shown promising results by including a stage where they continue pretraining Llama 2 7B and Llama 2 13B models on monolingual data covering their target languages. We think including such a stage before our SFT stage is worth considering in our future models.
- Optimization of model merging: Our experiments with checkpoint merging (Section 2.5) were extremely sparse. In the future, we would also like to evaluate SLERP and linear interpolation in comparable settings and a broader range of possible combined models (checkpoints from a single run vs. checkpoints across different run branches).

6 Conclusion

In this paper, we presented the CUNI submissions for the WMT24 General Translation task and the Translation into Low-Resource Languages of Spain task. Our primary focus was on using small open-source language models for various language pairs and providing comparisons with our systems from previous years.

The CUNI-MH system for English-to-Czech translation, based on Mistral 7B, showed promising results, possibly because of its CPO stage which led to a significant improvement of COMET and

Dataset	COMET22	COMETKIWI22	BLEU
WMT22	84.24	78.21	24.30
WMT23	75.33	74.81	21.63
WMT23-para	75.33	74.81	25.89

Table 9: CUNI-DS's segment-level scores for the first stage (en2cs training and en2cs evaluation) across different test datasets.

Dataset	COMET22	COMETKIWI22	BLEU
WMT22	85.81	80.97	24.45
WMT23	85.89	81.02	22.30
WMT23-para	72.27	78.21	21.63

Table 10: CUNI-DS's segment-level scores for the second stage (en2ru fine-tuning and en2ru evaluation) across different test datasets.

Model	COMET	BLEU	chrF
Apertium			
Aragonese*	0.788	65.3	82.0
Aranese	0.623	37.8	59.9
Asturian	0.652	16.9	50.6
Command-R 4-bit			
Aragonese	0.702	15.9	49.5
Aranese	0.576	4.5	33.3
Asturian	0.680	14.5	46.7
Aya-23			
Aragonese	0.685	12.4	46.7
Aranese	0.535	4.1	31.8
Asturian	0.645	9.0	40.3

Table 11: Scores of the baseline models on FLORES+ dev set in translation from Spanish into the given language. We note that the Aragonese part of the test set was created by post-editing Apertium translation, which is marked by the asterisk.

COMETKIWI scores, surpassing our previous systems. The model weights are available on Huggingface[12].

Our other submissions explored various techniques, such as transfer learning (CUNI-DS on en2ru), adaptation from speech translation (CUNI-NL on en2de) and creation of synthetic data using backtranslation to evaluate the feasibility of using LLMs for low-resource languages in the Translation into Low-Resource Languages of Spain task.

[12] https://huggingface.co/wmt24-cuni/CUNI-MH

Model	COMET	BLEU	chrF
Command-R 4-bit (240)			
Aragonese	0.779	37.9	69.7
Aranese	0.634	33.1	57.4
Asturian	0.699	15.3	49.0
Aya-23 (5000)			
Aragonese	0.780	39.0	70.2
Aranese	0.632	35.0	58.1
Asturian	0.686	15.2	48.8

Table 12: Scores of the fine-tuned models on FLORES+ dev set in translation from Spanish into the given language. Number of fine-tuning steps in the parentheses.

7 Acknowledgments

This work was partially supported by the Grant Agency of Charles University in Prague (GAUK 244523), GAČR EXPRO grant LUSyD (GX20-16819X), TAČR grant EdUKate (TQ01000458) and Czech Ministry of Industry and Trade grant CEDMO 2.0 NPO (NPO/1).

It has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

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

References

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech and Language*, 66:101155.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Mikel L. Forcada and Francis M. Tyers. 2016. Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. Pilar.

Petr Gebauer, Ondřej Bojar, Vojtěch Švandelík, and Martin Popel. 2021. CUNI systems in WMT21: Revisiting backtranslation techniques for English-Czech NMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 123–129, Online. Association for Computational Linguistics.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*.

Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. Knowledge is a region in weight space for fine-tuned language models.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

Josef Jon and Ondřej Bojar. 2023. Breeding machine translations: Evolutionary approach to survive and thrive in the world of automated evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2191–2212, Toronto, Canada. Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2022. CUNI-bergamot submission at WMT22 general translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 280–289, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI at WMT23 general translation task: MT and a genetic algorithm. In *Proceedings of the Eighth Conference on Machine Translation*, pages 119–127, Singapore. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: the LLM era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondřej Bojar. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *arXiv:2007.03006*.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. 2023. Loftq: Lora-fine-tuning-aware quantization for large language models.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. ArXiv:1711.05101 [cs, math].

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

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.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110:43–70.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

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.

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.

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, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

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

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '85, page 245–254, New York, NY, USA. Association for Computing Machinery.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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

Joachim Utans. 1996. Weight averaging for neural networks and local resampling schemes. In *AAAI-96 Workshop on Integrating Multiple Learned Models*, page 133–138. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

A CUNI-MH Model Prompt Template and Packing

We used the following prompt template for the model, inspired by the one used in Alpaca (Taori et al., 2023):

```
### Instruction:
Translate Input from English to Czech
### Glossary:

### Previous text:

### Input:
{source_text}
### Response:
{target_text}
```

The *Glossary* and *Previous text* sections were not used for the current task, so we left them empty. Since we trained only a single translation direction this time, the instruction remains constant.

Below is a shortened example of the packed[13] and tokenized training data, where <s> stands for the beginning of sequence token, </s> stands for the end of sequence token and \n stands for newline, the tokens are separated by spaces:

```
<s> ▁### ▁Inst ruction : \n Trans late ▁Input ▁from ▁English ▁to ▁Czech \n ### ▁Gl oss ary : \n \n ### ▁Pre vious ▁text : \n \n ### ▁Input : \n It ▁had ▁been ▁bad ▁enough , ▁calling ▁Brother ▁when ▁she ▁was ▁with ▁him . \n ### ▁Response : \n By lo ▁d ost ▁z lé ▁př iv ol at ▁Br atra , ▁k dy ž ▁byla ▁s ▁n ím . </s> <s> ▁### ▁Inst ruction : \n Trans late ▁Input ▁from ▁English ▁to ▁Czech \n ### ▁Gl oss ary : \n \n ### ▁Pre vious ▁text : \n \n ### ▁Input : \n To ▁do ▁it ▁now ? \n ### ▁Response : \n A le ▁te ď ? </s> <s> ▁### ▁Inst ruction : \n Trans late ▁Input ▁from ▁English ▁to ▁Czech \n ### ▁Gl oss ary : \n \n ### ▁Pre vious ▁text : \n \n ### ▁Input : \n Here ? \n ### ▁Response : \n T ady ? </s>
```

[13]The packing itself is implemented by TRL's ConstantLengthDataset, see `https://github.com/huggingface/trl/blob/e3fe28ee1a8bfab9739f849759c93d56776376e2/trl/trainer/utils.py#L431`

B CUNI-GA Model Prompt Template

We used the following prompt for context-aware translation in the Translation into Low-Resource Languages of Spain task, in order to make use of document-level context, while still keeping alignment on the sentence level, necessary for the evaluation:

```
We need to translate a single line from
conversation in Spanish into
{target_language}.  This is the
conversation: {src_context}

The start of the conversation is already
translated into English: {prev_context}
Translate the following line from
{src_lang} to {tgt_lang}.

Be very literal, and only translate the
content of the line, do not add any
explanations: {src_line}
```