Efficient Architectures For Low-Resource Machine Translation

Edoardo Signoroni
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czechia
e.signoroni@mail.muni.cz

Pavel Rychlý
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czechia
pary@fi.muni.cz

Ruggero Signoroni
Department of Information Engineering
University of Brescia
Via Branze 38, Brescia, Italy
r.signoroni001@studenti.unibs.it

Abstract

Low-resource Neural Machine Translation is highly sensitive to hyperparameters and needs careful tuning to achieve the best results with small amounts of training data. We focus on exploring the impact of changes in the Transformer architecture on downstream translation quality, and propose a metric to score the computational efficiency of such changes. By experimenting on English-Akkadian, German-Lower Sorbian, English-Italian, and English-Manipuri, we confirm previous finding in low-resource machine translation optimization, and show that smaller and more parameter-efficient models can achieve the same translation quality of larger and unwieldy ones at a fraction of the computational cost. Optimized models have around 95% less parameters, while dropping only up to 14.8% ChrF. We compile a list of optimal ranges for each hyperparameter.

1 Introduction

Neural machine translation (NMT) has done massive progress in high-resource conditions, due to the performance of models based on encoder-decoder architectures, such as the Transformer (Vaswani et al., 2017). Often, this progress did not trickle down to low or extremely low-resource languages, due to the huge requirements in terms of available training data and computational resources (Ranathunga et al., 2023). Default settings and assumptions for high-resource scenarios, such as the correlation of model size and performance, are not true in a low-resource one. While some attempts are being done to adapt and prompt large language models (LLMs) for low-resource machine translation (Guo et al., 2024; Lu et al., 2024; Merx et al., 2024; Aycock et al., 2025; Joshi et al., 2025; Khade et al., 2025), handling them is not always feasible nor convenient. Suitable hardware may not be available to deploy or train sufficiently large models. Moreover, even if capable hardware is available, using LLMs may still be a suboptimal choice, since it is much harder to freely fit the model and its architecture to the scarce data. Moreover, as Petrov et al. (2023) shows, LLMs' own pre-trained tokenizers are biased against low-resource languages due to their low share of training data. Considering that even state-of-the-art LLMs underperform standard NMT (Robinson et al., 2023), employing them when smaller options are available is inefficient.

Training a Transformer in these settings remains a challenging task, and one that requires careful hyperparameter tuning (Popel and Bojar, 2018). However, if done correctly, it can lead to well-performing and competitive models (van Biljon et al., 2020; Araabi and Monz, 2020). Most of the work regarding low-resource machine translation focuses on several techniques, such as fine-tuning, or transfer learning (Ranathunga et al., 2023). Research on scaling and optimizing machine translation has mainly been done in a high-resource setting (Ghorbani et al., 2022), or on other aspects of training (Sennrich and Zhang, 2019; Araabi and Monz, 2020; Signoroni and Rychlý, 2024).

Following the finding that not only size, but also shape of the Transformer influences downstream performance (Tay et al., 2022), our work aims to broaden the understanding of the scaling of machine translation in low-resource settings by experimenting with four key components in the architecture of the Transformer model: encoder layers, decoder layers, embedding size, and feedforward dimension. We conduct experiments on one simulated low-resource pair, and three true low-resource

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 39–64, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-006

pairs, to explore the impact of each hyperparameter on the downstream translation task. We propose a novel **P**arameter **I**ncrease **E**fficiency **Score** (PIES) to measure the efficiency of changing the configuration of the model, and to find the most parameter-efficient combinations for each dataset. We compile a list of empirically-found optimal ranges for each hyperparameter to inform future exploration and training of low-resource machine translation models.

We confirm that in low-resource conditions the Transformer is highly susceptible to hyperparameter variation. We also find that smaller models can perform as well as much bigger models, at just a tiny fraction of the computational cost.[1]

2 Related Work

Our work intersects previous studies on Transformer and Machine Translation scaling laws and optimization on both high and low-resource languages.

2.1 Scaling Laws and Optimization

Works tackled the challenge of finding empirical scaling laws that govern neural language model scaling, considering model, computational, or dataset size.

Tay et al. (2022) conduct extensive experiments involving over 200 Transformer configurations considering both upstream and several downstream tasks (though, crucially, not machine translation). They find that model shape, and not only size (Kaplan et al., 2020), strongly influences downstream performance. They also find that scaling laws change substantially when considering metrics on actual downstream fine-tuning. Notably, they show that scaling strategies differ at different compute regions, and thus finding strategies at small scale might not necessarily transfer or generalize to higher compute regions.

Some work has also been conducted for machine translation.

Ghorbani et al. (2022) explore scaling laws for machine translation on a high-resource English-German dataset. Their results indicate that the scaling behavior is largely determined by the total capacity of the model, and its allocation between the encoder and the decoder. Moreover, they suggest that scaling behavior of encoder-decoder NMT models is predictable, but the scaling laws might vary depending on the particular architecture or task.

Gordon et al. (2021) study the predictability of MT system performance as parameters/data increase. They train many Transformers of various sizes randomly selected subsets of data for Russian-English, German-English, and Chinese-English. Crucially, they find that extending their previous experiments to datasets smaller than 50MB, using 0.05% - 0.0125% of the data, the data scaling power law breaks down, indicating the impossibility of extrapolating extremely low-resource results to medium and high-resource data regimes.

Some research (Hsu et al., 2020; Kasai et al., 2021; Berard et al., 2021) has also departed from the convention of using balanced encoder and decoders, resulting in "deep encoder, shallow decoder" models that can speed up inference while maintaining a similar translation performance.

2.2 Optimization for Low-Resource Settings

Some studies have also been done on optimizing NMT for low-resource scenarios.

Sennrich and Zhang (2019) find that best practices differ between high-resource and low-resource MT and that the latter is highly sensitive to hyperparameters by training RNNs with different techniques and hyperparameters on a simulated English-German, and a true Korean-English low-resource dataset.

Araabi and Monz (2020) trains Transformers for a diverse set of true and simulated low-resource pairs to find that a proper combination of Transformer configurations results in substantial improvements over a Transformer system with default settings. For example, they observe that a shallower Transformer combined with a smaller feed-forward layer dimension and two attention heads is more effective.

van Biljon et al. (2020) experiment with different Transformer configurations on the translation of three low-resource languages, showing that medium (6 total layers) and shallow (2 total layers) perform better than the canonical configuration of 6 encoder and 6 decoder layers.

[1]Full results and code is available at https://github.com/edoardosignoroni/eff_archs_lowre

3 Methodology

This section describes the dataset we tested on (Section 3.1)[2]. It then reports the training framework and the hyperparameters we used (Section 3.2). Next, it explains our proposed efficiency metric (Section 3.3). And finally, it outlines our experimental setup (Section 3.4).

3.1 Datasets

Our experiments are carried out on publicly available low-resource datasets, and one simulated low-resource dataset retrieved from OPUS (Tiedemann, 2009). The datasets involve both high-resource languages (English, German, Italian), and a selection of under-resourced languages (Akkadian, Lower Sorbian, Manipuri). The datasets have between 21k and 50k sentence pairs, thus can be considered as extremely low-resource (Ranathunga et al., 2023). Their content is from different domains, mainly news and Wikipedia text, except for Akkadian, which is mostly assorted fragments of cuneiform texts. The low-resource datasets have their own validation and test splits, while for the simulated English-Italian dataset we use the *dev* and *devtest* splits from the Flores-200 benchmark corpus (Goyal et al., 2022). The datasets are summarized in Table 1.[3]

3.2 Hyperparameters and Training

After tokenizing the data using BPE (Sennrich et al., 2016), as implemented in SentencePiece (Kudo and Richardson, 2018), we learn separated vocabularies for source and target with a size of 4k items, without a frequency threshold.

We train Transformers (Vaswani et al., 2017) with Fairseq (Ott et al., 2019) until BLEU score on validation does not increase for 20 consecutive epochs or until 50000 updates. As our baseline, we chose a *small* model that performed sufficiently well in previous experiments for all pairs. Its architecture and training hyperparameters are given in Table 2. We share embeddings between the encoder and the decoder. Each model is trained on a single Nvidia A40 or A100 GPU.

During the experiments, we focus on tuning the architecture of the model by changing the number of encoder and decoder layers, the size of the embeddings, and the feed forward dimension. We leave all other hyperparameters unchanged. We leave the number of heads at 2, following Araabi and Monz (2020).

We will refer to the models with the following naming scheme: *enc_dec_embs_ffw_heads*. E.g. our baseline model may be referred as *4_4_256_1024_2*.

3.3 Efficiency Score

To evaluate the efficiency of the models, we introduce a **P**arameter **I**ncrease **E**fficiency **S**core, or **PIES**, computed as follows:

$$PIES = \frac{size/score}{10^6}$$

where score means a machine translation metric such as COMET, ChrF, or BLEU, and size means the total number of parameters of the model. Thus, PIES is computed as the ratio between the size of the model in number of parameters and the machine translation score it achieved, divided by 1 million. This is an easily interpretable and straightforward metric that gives the millions of parameters needed for each score point. A lower value denotes a more efficient system.

We compute the total number of parameters for each model as follows:

$$params = (2 \times E \times V) + (4 \times E^2 + 2 \times E \times F + 9 \times E + F) \times enc + (8 \times E^2 + 2 \times E \times F + 15 \times E + F) \times dec$$

where E is the size of the embeddings, V the number of items in the vocabulary, F is the feed-forward dimension, and *enc/dec* is the number of layers in the encoder/decoder, respectively.

To obtain the score for each model after training, we generate test set translations for each model and obtain sentence-level BLEU (Papineni et al., 2002), ChrF (Popović, 2015), ChrF++ (Popović, 2017), and COMET (Rei et al., 2020) scores as implemented in Hugging Face `evaluate` library.[4] We employ bootstrap evaluation on 200 batches of 400 test sentences to obtain the final scores.

Mathur et al. (2020) argue for the retirement of BLEU in favour of ChrF++. Sai B et al. (2023)

[2]Appendix A provides more information about the languages involved.

[3]We use a simple Python script to split the tokenized data at the newline character and the whitespace and then return the length of the resulting lists to obtain the number of lines and tokens for each pair.

[4]Scores for metrics other than ChrF are available in the Appendices (Tables 17-23) and in the GitHub repository.

Languages	Abbreviation	Dataset	N. of Pairs	Src Tokens	Tgt Tokens
English-Akkadian	eng-akk	EvaCun 2023	45,269	1,177,138	630,535
German-Lower Sorbian	deu-dsb	WMT22 Low-res shared Task	40,194	1,064,087	1,032,701
English-Italian	eng-ita	WikiMatrix Random Selection	50,000	1,571,843	1,723,391
English-Manipuri	eng-mni	WMT23 Indic Shared Task	21,287	748,407	715,548

Table 1: **Summary of the datasets in our experiments.** The columns report the languages in the dataset, its original source, and the size of the training split in number of tokens and sentence pairs.

Parameters	
vocabulary size	4,000
encoder layers	4
decoder layers	4
enc/dec embedding dim	256
enc/dec feed forward dim	1,024
enc/dec attention heads	2
optimizer	adam
adam betas	0.9, 0.98
learning rate	1e-4
warmup updates	5,000
dropout	0.1
label smoothing	0.1
max tokens	16,000

Table 2: **Hyperparameters for our baseline model.** For the other models in our experiments, we change only the number of layers, the size of the embeddings, and the feed forward dimension.

finds that ChrF++ performs the best among overlap metrics for a selection of Indic languages. The results of recent WMT Metrics shared tasks (Freitag et al., 2022) demonstrate that learned neural metrics are the most optimal. Among these, COMET is the current state-of-the-art, and is widely employed in machine translation studies. However, pretrained neural metrics are unreliable for unseen languages, especially under-resourced ones. Works such as the ones by Sai B et al. (2023) and Wang et al. (2024) show that fine-tuned COMET models perform better for specific sets of low-resource languages, than baseline models. For these reasons we chose ChrF as the metric of reference in both our observations and PIES.

By computing Pearson's r between PIES and ChrF score on the aggregate results of our experiments, we obtain r=-0.543, indicating a negative correlation between PIES and translation quality: a lower PIES corresponds to a higher ChrF.

3.4 Experiments

Our aim is to investigate efficient architectures for low-resource machine translation models by tuning hyperparameters such as encoder and decoder layers, embeddings and feed forward dimension. We fix all other training hyperparameters to values found to be optimal or close to optimal in previous and preliminary experiments on the same data (Signoroni and Rychlý, 2024).

3.4.1 Experiment 1: Change One, Fix All

Hyperparameters	
encoder layers	2, **4**, 6, 8, 12, 16, 24, 32
decoder layers	2, **4**, 6, 8, 12, 16, 24, 32
embedding dimension	**256**, 512, 1024, 2048, 4096
feed forward dimension	256, 512, **1024**, 2048, 4096

Table 3: **Values for each hyperparameter tried in Experiment 1.** Baseline values are in bold.

Size	Hyperparameters	N. of Parameters
small	4_4_256_1024_2	8.5M < **9.4M** < 10M
base	6_6_512_2048_2	40M < **48M** < 53M
large	6_6_1024_4096_2	166M < **184M** < 203M

Table 4: **Baseline hyperparameters and sizes (in bold) for the models in Experiment 2.** We consider all possible architectures in a range of $\pm 10\%$ parameters from these baseline models.

Our first experiment focuses on changing only one hyperparameter at a time in the architecture of the model without controlling the total amount of parameters. We start from our baseline values of 4 encoder and decoder layers, embedding size of 256, and feedforward dimension of 1024, and change them one step at a time according to Table 3.

3.4.2 Experiment 2: Parameters Budget

In Experiment 2, we fix the number of parameters to $\pm 10\%$ of transformer small, base, and large and test all possible combinations of hyperparameters

that fall into the ranges given in Table 4. For each dataset, we test each possible configuration that falls within these ranges: 13 for *small* (counting the baseline 4_4_256_1024_2 model), 58 for *base*, and 60 for *large*, that is 131 combinations for dataset, for a total of 524 models. By allowing all possible combinations of hyperparmeters, we overcome one limitation of the previous setup, that is the chance of missing possible optimal configurations due to changing only one hyperparameter at a time.

4 Results

4.1 Experiment 1: Change One, Fix All

As we discuss the results of our experiments, recall that a lower PIES denotes a more efficient model.

In Experiment 1, we start from the baseline 4_4_256_1024_2 model and increase or decrease only one hyperparameter at a time, leaving all other unchanged. Table 5 summarizes the results of Experiment 1.

As expected, increasing the embedding size leads to the biggest increase in model size, since it scales quadratically with the amount of parameters. Conversely, all the other hyperparameters we considered scale linearly with the number of parameters, with feedforward dimension being the least impactful per unit. Increasing the number of decoder layers results in a slightly steeper rate of increase in parameters than adding more encoder layers.

In this experimental setup, we allow the model size to grow freely. We observe that for all datasets increasing embedding size to 2048 or 1024 leads to the best ChrF scores, but also to disproportionally big models, reaching 75M or 251M parameters. For all four datasets, it is sufficient to scale back both embedding size and feedforward dimension to 256 to obtain the most efficient configuration. These optimized models have between 91.6% and 97.5% parameters less than the best architectures according to ChrF, while losing 1%-13.7% of the translation performance. We argue this is a favourable trade-off, especially in a low-resource setting where it may be needed to train several models in sequence for techniques such as back-translation.

As a matter of comparison, we prompted two language models, mt5-small (300M parameters) (Xue et al., 2021) and Mistral Small (24B parameters)[5], for translation in a zero-shot scenario. mt5-small fails at translating between all pairs, even the high-resource English-Italian[6], reaching only 3.1 ChrF for *deu-dsb*. Consequently, using the model in this way is very inefficient, as reflected by hugely inflated PIES scores. The much bigger Mistral Small fares much better for English-Italian, achieving 59 ChrF, thus making it the best model overall in our experiment for this pair. Its efficiency is, however, debatable, as indicated by PIES scores in the hundreds and thousands. With respect of the low-resource pairs, Mistral Small fails for Akkadian and Manipuri, and performs poorly for Lower Sorbian.

We then trained a new mt5-small model for each pair by finetuning the original mt5-small on each dataset for 5 epochs. As expected, this increased ChrF and decreased PIES across all pairs. The best performance is for English-Italian, which reaches a ChrF of 44 points, comparable to other models trained from scratch. PIES is still quite high, however, at 6.8. Finetuning significantly helps mt5-small to make sense of unseen low-resource languages, especially with Lower Sorbian and Manipuri, but both performance and efficiency are well below models trained from scratch. This suggests that when only a small amount of data is available, training from scratch may still be the best choice, especially if the under-resourced languages in question are not in the training data of the language model to finetune or adapt.

Table 7 gives ChrF and PIES for two language models prompted for translation across the language pairs in our experiment.[7]

4.2 Experiment 2: Parameters Budget

In Experiment 2, we limit the number of parameters in three ranges, corresponding to the sizes of Transformer *small*, *base*, and *large* (Table 4). The higher number of combinations per dataset (131) allows for observations regarding some *average* trends in our results. To extrapolate optimal ranges for the hyperparameters and their interactions, we proceed in three steps: 1. we filter out all combinations with a ChrF<35; 2. among these, we keep only those with PIES<1; 3. we select ranges where the remaining combinations are optimal across the majority of the datasets. To generalize these ranges,

[5]https://mistral.ai/news/mistral-small-3-1

[6]English and Italian are known to mt5. English has the biggest share of the training data (5.67%), and Italian has the sixth biggest (2.63%) of the total amount.

[7]Refer to Appendix B for information about the prompting and finetuning of these models.

	eng-akk	deu-dsb	eng_wiki-ita_wiki	eng-mni
Best Model (ChrF)	**4_4_2048_1024_2**	**4_4_2048_1024_2**	**4_4_2048_1024_2**	**4_4_1024_1024_2**
ChrF	41.792	48.881	45.612	48.505
PIES	6.017	5.145	5.513	1.555
Num. Parameters	251M	251M	251M	75M
Best Model (PIES)	**4_4_256_256_2**	**4_4_256_256_2**	**4_4_256_256_2**	**4_4_256_256_2**
ChrF	39.681	44.527	45.156	41.844
PIES	0.158	0.141	0.139	0.150
Num. Parameters	6.3M	6.3M	6.3M	6.3M
Δ ChrF ↑	**-2.111**	**-4.354**	**-0.456**	**-6.661**
% of best	**-5.052%**	**-8.907%**	**-1.000%**	**-13.732%**
Δ PIES ↓	**-5.859**	**-5.004**	**-5.374**	**-1.405**
% of best	**-97.374%**	**-89.308%**	**-97.479%**	**-90.354%**
Δ Params ↓	**-245M**	**-245M**	**-245M**	**-69M**
% of best	**-97.507%**	**-97.490%**	**-97.507%**	**-91.600%**

Table 5: **Best models from Experiment 1 according to ChrF and PIES.** Below the model name (in the form *enc_dec_embs_ffw_heads*), we report ChrF, PIES, and size of the model. In the bottom part of the table, we report the differences in scores ($\Delta ChrF \uparrow$, $\Delta PIES \downarrow$) and size ($\Delta Params \downarrow$) between the best model for translation quality (highest ChrF) and the most efficient one (lowest PIES), both in absolute terms and as a percentage (% of best).

	eng-akk	deu-dsb	eng_wiki-ita_wiki	eng-mni
Best Model (ChrF)	**6_8_1024_2048_2**	**12_2_1024_4096_2**	**12_2_1024_4096_2**	**2_16_1024_256_2**
ChrF	43.393	51.569	47.890	49.882
PIES	3.673	3.741	4.029	3.217
Num. Parameters	176M	193M	193M	160M
Size range	large	large	large	large
Best Model (PIES)	**4_6_256_512_2**	**6_2_256_1024_2**	**6_2_256_1024_2**	**4_8_256_256_2**
ChrF	38.811	44.001	45.347	44.483
PIES	0.229	0.202	0.196	0.200
Num. Parameters	8.9M	8.9M	8.9M	8.9M
Size range	small	small	small	small
Δ ChrF ↑	**-4.582**	**-7.568**	**-2.543**	**-5.339**
% of best	**-10.559%**	**-14.790%**	**-5.310%**	**-10.823%**
Δ PIES ↓	**-3.444**	**-3.539**	**-3.833**	**-3.017**
% of best	**-93.375%**	**-94.600%**	**-95.135%**	**-93.783%**
Δ Params ↓	**-167M**	**-184M**	**-184M**	**-151M**
% of best	**-94.943%**	**-95.389%**	**-95.389%**	**-94.438%**

Table 6: **Best models from Experiment 2 according to ChrF and PIES.** Below the model name (in the form *enc_dec_embs_ffw_heads*), we report ChrF, PIES, and size of the model. In the bottom part of the table, we report the differences in scores ($\Delta ChrF \uparrow$, $\Delta PIES \downarrow$) and size ($\Delta Params \downarrow$) between the best model for translation quality (highest ChrF) and the most efficient one (lowest PIES), both in absolute terms and as a percentage (% of best).

we filter all combinations according to the values we found, starting with the most impactful ones. Table 8 summarizes these findings, and Table 9 reports the possible optimal configurations we found.

At a glance, it can be observed that pruning models that are either too deep or too imbalanced in terms of encoder and decoder layers leads, on average, to better ChrF and PIES, and greatly reduces the size of the models. Limiting embedding size may reduce quality, however this can be circumvented by selecting balanced architectures according to other criteria. We can also observe that setting the feedforward dimension in the suggested range, when taken in relation to the number of encoder and decoder layers, slightly increases both ChrF and PIES. Further experiments may ad-

Model	Metric	eng-akk	deu-dsb	eng_wiki-ita_wiki	eng-mni
mt5-small	**ChrF**	0.070	3.101	1.584	0.004
(300M)	**PIES**	*4271.370*	*96.729*	*189.406*	*82101.976*
mt5-small-finetuned	**ChrF**	7.136	32.681	44.033	25.687
(300M)	**PIES**	*42.043*	*9.180*	*6.813*	*11.679*
mistral-small	**ChrF**	0.358	18.332	59.086	4.966
(24B)	**PIES**	*67039.379*	*1309.188*	*406.190*	*4832.926*

Table 7: **ChrF and PIES for zero-shot mt5-small, finetuned mt5-small, and zero-shot Mistral-Small.**

Filter	Low	High	Remaining Configurations	Avg ChrF	Avg PIES	Avg Size (in M)
Initial	-	-	524	33.93	5.06	102M
layer_sum	6	18	244	42.137	2.087	84.2
embs	256	512	148	40.659	0.898	31.8
enc_layers	2	12	136	41.491	0.871	31.7
dec_layers	2	12	132	41.653	0.806	31.3
layer_diff	-6	8	116	41.999	0.771	30.8
ffw	256	2048	92	42.163	0.675	27.3
enc/dec	0.25	3	92	42.163	0.675	27.3
embs/ffw	0.125	2	92	42.163	0.675	27.3
embs/enc	21.333	256	92	42.163	0.675	27.3
embs/dec	32	128	92	42.163	0.675	27.3
embs/num_layers	16	51.2	92	42.163	0.675	27.3
ffw/dec	21.333	1024	92	42.163	0.675	27.3
ffw/enc	32	1024	88	42.3	0.695	28.1

Table 8: **Optimal ranges of the hyperparameters and their interactions.** The first two columns give minimum and maximum values for each one. The other report the remaining configurations after filtering the possible combinations, and their average ChrF, PIES, and size in millions of parameters.

dress the impact of this particular relation. In summary, optimal architectures should have a limited number of layers (6 to 18), which must be not too unbalanced on either encoder or decoder side (-6 to 8 encoder - decoder difference; 0.125 to 3 encoder/decoder ratio). Embedding size should be kept on the smaller side (256 to 512), while feed-forward dimension can be bigger (up to 2048).

If we apply these guidelines to the full results, without prior filtering for ChrF and PIES, we are left with the 22 combinations in Table 9, out of the initial 131 per dataset. Apart from four, all have an average ChrF > 40, and six have an average PIES > 1. All of these are of *base* size. There is a noticeable gap between PIES for this size bracket and *small*, which features the best efficiency scores.

Table 6 reports the best models and scores in Experiment 2. All the best model according to ChrF are in the *large* range, whereas the most efficient ones according to PIES are in the *small* bracket. For two datasets, *deu-dsb* and *eng_wiki-ita_wiki*, the best ChrF model is the same (12_2_1024_4096_2). The best ChrF model for *eng-mni* is quite peculiar:

Model	Avg ChrF	Avg PIES	Size (in M)
6_2_256_1024_2	42.757	0.209	8.9
4_6_256_512_2	42.062	0.212	8.9
4_4_256_1024_2	42.836	0.221	9.4
8_4_256_512_2	41.845	0.226	9.4
6_8_256_256_2	42.470	0.229	9.7
6_6_256_512_2	42.859	0.233	10.0
4_8_256_256_2	39.176	0.242	8.9
2_8_256_512_2	40.237	0.247	9.4
2_6_256_1024_2	40.762	0.254	9.9
8_6_256_256_2	36.428	0.316	9.2
8_6_512_1024_2	45.171	0.885	39.8
6_4_512_2048_2	44.738	0.894	39.8
4_6_512_2048_2	45.423	0.925	41.9
6_8_512_1024_2	44.926	0.937	41.9
12_4_512_1024_2	43.269	0.978	41.9
2_8_512_2048_2	44.249	0.998	44.0
8_4_512_2048_2	45.516	1.015	46.1
6_6_512_2048_2	44.920	1.077	48.2
12_6_512_1024_2	43.353	1.122	48.3
6_12_512_512_2	40.087	1.258	45.1
6_12_512_256_2	38.743	1.316	40.4
8_8_512_1024_2	38.762	1.484	46.2

Table 9: **Optimal model configurations** and their size, with ChrF and PIES averaged over all four datasets.

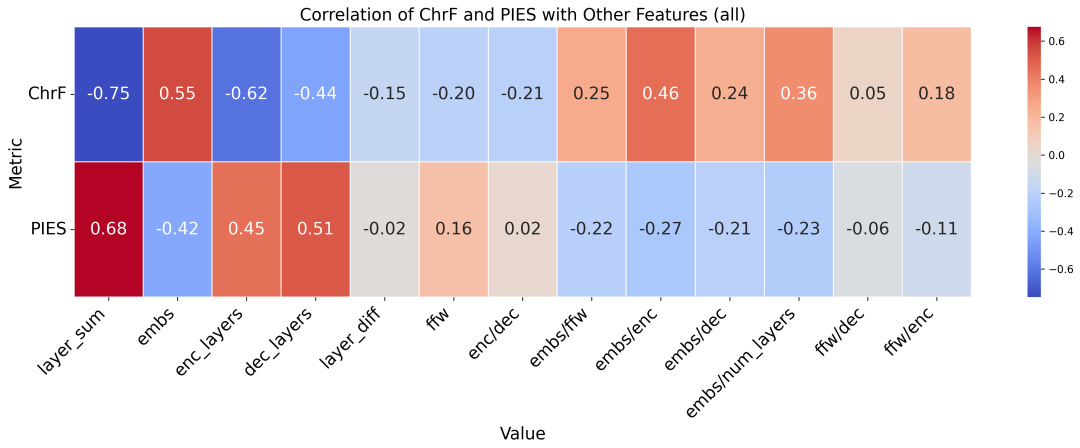

Figure 1: **Correlation matrix of ChrF and PIES** with other hyperparameters and their interactions. A lower PIES is better.

it has just 2 encoder layers, 16 decoder layers, an embedding size of 1024, and a narrow feedforward of just 256. Again we see decrements in ChrF between 5.3% and 14.8%, against a 95% reduction in number of parameters. While bigger models may in principle achieve a slightly higher ChrF, this comes at the cost of efficiency. We argue that in a low-resource scenario, when both data and hardware are scarce, the increased computational cost needed to find and train the optimal model in this size range is not well spent. Smaller models can achieve a comparable, or almost comparable translation performance, at just a fraction of the cost. This is also true for exploratory runs, or intermediate stages of development, such as systems for backtranslation.

From Figure 1, reporting the correlation between ChrF and PIES and all the features in our study, we can point out some interesting observations. The number of layers is the most impactful hyperparameter for both ChrF and PIES. Deeper architectures lose both in terms of quality and efficiency. It follows that the number of encoder and decoder layers impacts the metrics, with changes in the encoder slightly more important for ChrF, and changes in the decoder a bit more impactful on PIES. Embedding dimension is relevant for both metrics. Feedforward appears to have a lesser impact. According to these findings, **balancing embeddings dimension, the number of layers** in the encoder and the decoder, **and their interactions is key** to an efficient model with good translation quality.

5 Conclusions

In this paper, we explored scaling and optimizing the Transformer architecture for low-resource machine translation by experimenting with several hundred configurations over four language pairs.

We confirm previous findings that the Transformer, and low-resource NMT in general, is highly sensitive to hyperparameters in low-resource conditions, and that standard settings are not optimal. We observe some trends and interactions between the number of encoder and decoder layers, embedding size, feedforward dimension, and the quality of the translation.

We propose PIES as a novel metric to measure the efficiency of changing a model's architecture, and use it to show that increasing model size is not always the optimal choice, since smaller and balanced models can reach a comparable performance for a fraction of the computational cost. We also outline some empirical findings and guidelines regarding the optimal hyperparameter ranges that result in more efficient low-resource machine translation models.

Limitations

The main limitations of our experiments are the following. First, the dataset selection, while trying to be diverse both in terms of typology (Germanic, Slavic, Romance, Tibeto-Burman, Semitic) and writing system (Latin, Bengali, Cuneiform), is only a tiny fraction of the world's 7000+ languages. If we include, also historical ones, such is the case with Akkadian, the number grows even more. We acknowledge that this fact may hinder

generalization, and to avoid even more grid search and computation, we attempted to gain as much information as possible from these datasets. We leave to future work to test our intuitions on a wider range of languages.

Second, we could not perform a systematic qualitative analysis on the outputs of the models, and had to rely on automated metrics to score the translations. This comes with another set of problems altogether, that is out of the scope of this paper to discuss. This is also relevant for PIES, which in its present iteration is closely correlated with the translation metric. In the future, we plan to extend it to account for multiple metrics, and to consider also train and inference times, and environmental concerns. For now, it is only as good as the translation metric chosen to compute it.

Lastly, we are aware that testing all possible combinations, across all hyperparameters, is a monumental task that evades the scope of just one paper. We focused on four specific architecture hyperparameters and their interactions. Other possible optimal configurations, that may need other changes in training hyperparameters (e. g. learning rate, dropout, etc.) to work best are left to future work. The same can be said for all LLMs, for which architecture cannot be modified as freely, with one led to employ different approaches such as fine-tuning and prompting techniques. These fall outside the scope of this paper and are left to future work.

Ethical Considerations

We did not collect any new data for these experiments, as we used publicly available dataset or parts thereof. The systems we trained are not intended to be deployed or used in any actual translation scenario, in such a case, they will incur in biases, errors, and issues common to this kind of NLP models, and as such they should be used with care. We are also aware of the environmental cost of training language models and tried our best to avoid grid search all the while getting a meaningful picture of the topic at hand.

Following (Lacoste et al., 2019), we estimate that our experiments lasted 4200 GPU hours on a private infrastructure with a carbon efficiency of 0.59 $kgCO_2eq/kWh$ for a total emissions of 708 $kgCO_2eq$.

Acknowledgments

We would like to thank the reviewers for their useful comments. This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062.

References

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? In *The Thirteenth International Conference on Learning Representations*.

Alexandre Berard, Dain Lee, Stephane Clinchant, Kweonwoo Jung, and Vassilina Nikoulina. 2021. Efficient inference for multilingual neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8563–8583, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2022. Scaling laws for neural machine translation. In *International Conference on Learning Representations*.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán,

and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Gai Gutherz, Shai Gordin, Luis Sáenz, Omer Levy, and Jonathan Berant. 2023. Translating Akkadian to English with neural machine translation. *PNAS Nexus*, 2(5):pgad096.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Yi-Te Hsu, Sarthak Garg, Yi-Hsiu Liao, and Ilya Chatsviorkin. 2020. Efficient inference for neural machine translation. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 48–53, Online. Association for Computational Linguistics.

Rudali Huidrom, Yves Lepage, and Khogendra Khomdram. 2021. EM corpus: a comparable corpus for a less-resourced language pair Manipuri-English. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 60–67, Online (Virtual Mode). INCOMA Ltd.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2025. Challenges in adapting multilingual LLMs to low-resource languages using LoRA PEFT tuning. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 217–222, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-English machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Marion Weller-di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

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

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt,

Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

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.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

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.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

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

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychlý. 2024. Better low-resource machine translation with smaller vocabularies. In *Text, Speech, and Dialogue*, pages 184–195, Cham. Springer Nature Switzerland.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations*.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

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.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayede, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed a nd Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Languages

Lower Sorbian (*"Dolnoserbšćina"*) is a West Slavic language predominantly spoken in eastern Germany by approximately 7,000 native speakers. Most of these speakers are from older generations, making the language critically endangered. Written in Latin script with additional diacritics, Lower Sorbian features six grammatical cases and a dual number system for nouns, pronouns, adjectives, and verbs. It does not employ articles. The dataset for our experiments was compiled by the Witaj Sprachzentrum [8] (Witaj Language Centre) (Weller-di Marco and Fraser, 2022).

Manipuri (*"Meiteilon"*) is a Tibeto-Burman language recognized as one of the official languages in the Indian state of Manipur and at the national level. It is spoken by approximately 1.8 million native speakers, primarily the Meitei people, both in Manipur and neighboring regions. UNESCO classifies Manipuri as "vulnerable." The language exhibits extensive suffixation with limited prefixation and follows an SVO word order. Other linguistic characteristics include agglutinative verb morphology, tone, a lack of grammatical person, number, and gender distinctions, and a focus on aspect rather than tense (Pal et al., 2023). Manipuri is written using several scripts, including the Meitei and Bengali scripts, with the latter being used for all the Manipuri data in our experiments. The Latin script is also employed. The dataset is a modified version (Pal et al., 2023) based on previous work by Haddow and Kirefu (2020), Laitonjam and Ranbir Singh (2021), and Huidrom et al. (2021). Each segment of the data set contains mainly news and other informational texts.

Akkadian, an extinct East Semitic language, was spoken in ancient Mesopotamia from the third millennium BCE until the 1st century CE. It utilized the cuneiform script, a logophonetic writing system in which symbols could serve as logograms, determinatives, or phonograms/syllabograms, each with a distinct interpretation. Akkadian is a fusional language with grammatical case and employs a root-based consonantal system. The dataset used in our study is derived from portions of the ORACC corpus [9] and mainly comprises Neo-Assyrian royal inscriptions and administrative correspondence. The stylistic variation between genres poses challenges for NLP (Gutherz et al., 2023). Additionally, because of the medium of preservation (clay tablets), the data is often incomplete, with truncated sentences.

[8] https://www.witaj-sprachzentrum.de/

[9] https://oracc.museum.upenn.edu/index.html

B Large Language Models' Prompting and Finetuning

mt5-small was prompted and finetuned using the HuggingFace Transfomers library, while Mistral Small was prompted with Ollama. For mt5-small, we used the prompt 'Translate SRC to TGT: SRC_EXAMPLE'. For fine-tuning, we restructured the data in a similar way, by giving the prompt, the source line, as input, and the target line as label. For Mistral Small, we opted for a somewhat more complex 'Translate the following text from SRC to TGT. Write only the translation. SRC: SRC_EXAMPLE TGT: '

C Tables and Charts

Below we report the results of our experiments. Tables 10 and 11 summarize the ChrF and PIES scores for each pair and size bracket of Experiment 2. Figures and plot the trend of ChrF when modifying each hyperparameter for each pair. Figures 4 and 5 show the counts of optimal configurations (ChrF>35 and PIES<1) for each hyperparameter across all datasets. Figure 6 report the average ChrF and PIES for the configurations in the optimal range for each language pair and size bracket. Finally, Tables 12 to 18 contain the BLEU, ChrF, and COMET scores for all combinations in Experiment 1 and 2.

tgt	size_tag	min	max	median	mean
akk	base	4.901	42.568	19.969	23.426
akk	large	6.088	43.394	30.677	26.093
akk	small	14.408	39.211	35.026	32.287
dsb	base	3.017	48.324	41.664	32.098
dsb	large	2.898	51.569	47.095	36.073
dsb	small	37.982	44.329	43.008	42.741
ita_wiki	base	2.928	46.282	43.531	33.336
ita_wiki	large	1.599	47.890	44.698	34.910
ita_wiki	small	42.504	45.347	44.919	44.708
mni	base	7.917	48.299	43.241	38.805
mni	large	7.733	49.883	44.606	40.643
mni	small	35.996	45.960	43.443	42.614
all	base	2.928	48.324	38.619	31.915
all	large	1.599	51.569	41.837	34.430
all	small	14.408	45.960	42.987	40.588

Table 10: Minimum, Maximum, Average, and Median ChrF values by language and size bracket in Experiment 2

tgt	size_tag	min	max	median	mean
akk	base	0.935	9.100	2.196	2.907
akk	large	3.673	29.323	5.951	10.493
akk	small	0.229	0.636	0.284	0.315
dsb	base	0.848	14.655	1.076	2.614
dsb	large	3.131	59.464	3.950	9.467
dsb	small	0.202	0.255	0.219	0.221
ita_wiki	base	0.861	16.206	1.067	2.378
ita_wiki	large	3.346	101.225	4.204	9.545
ita_wiki	small	0.196	0.228	0.210	0.211
mni	base	0.836	5.048	1.031	1.395
mni	large	3.218	22.284	3.953	5.422
mni	small	0.200	0.269	0.217	0.223
all	base	0.836	16.206	1.165	2.322
all	large	3.131	101.225	4.397	8.732
all	small	0.196	0.636	0.222	0.243

Table 11: Minimum, Maximum, Average, and Median PIES values by language and size bracket in Experiment 2

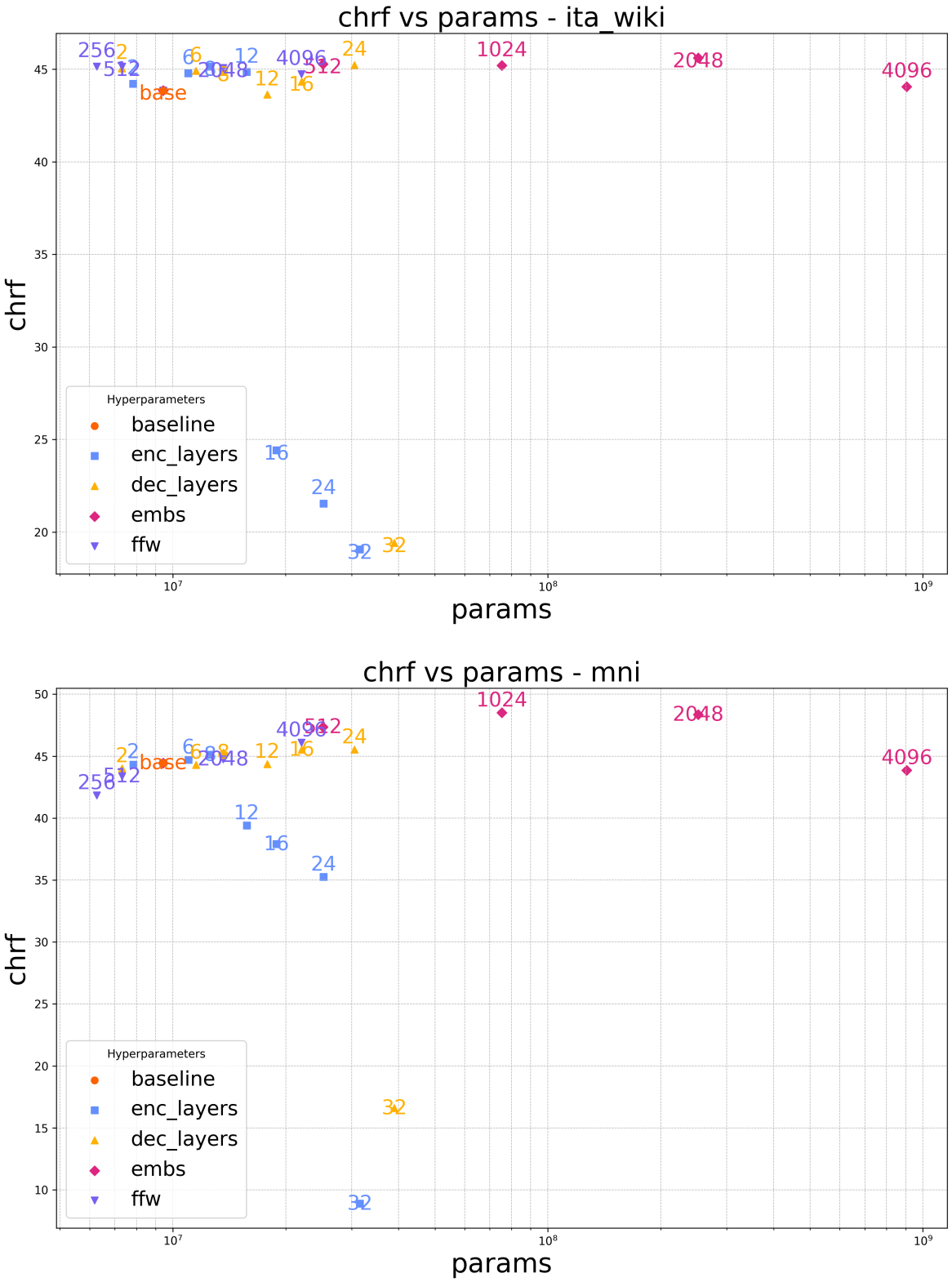

Figure 2: Charts plotting ChrF for each hyperparameter against the increase in number of parameters.

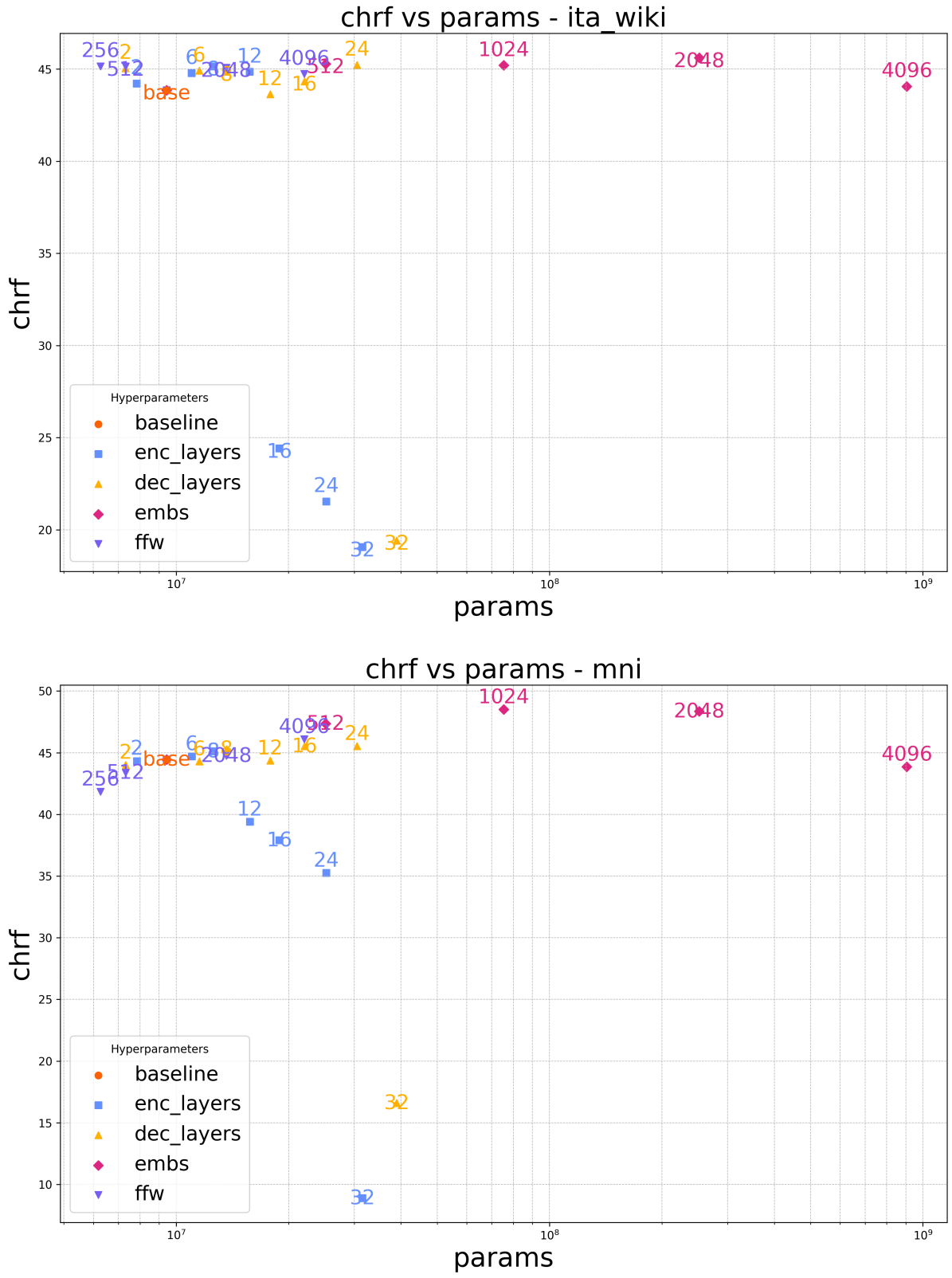

Figure 3: Charts plotting ChrF for each hyperparameter against the increase in number of parameters. (Continued)

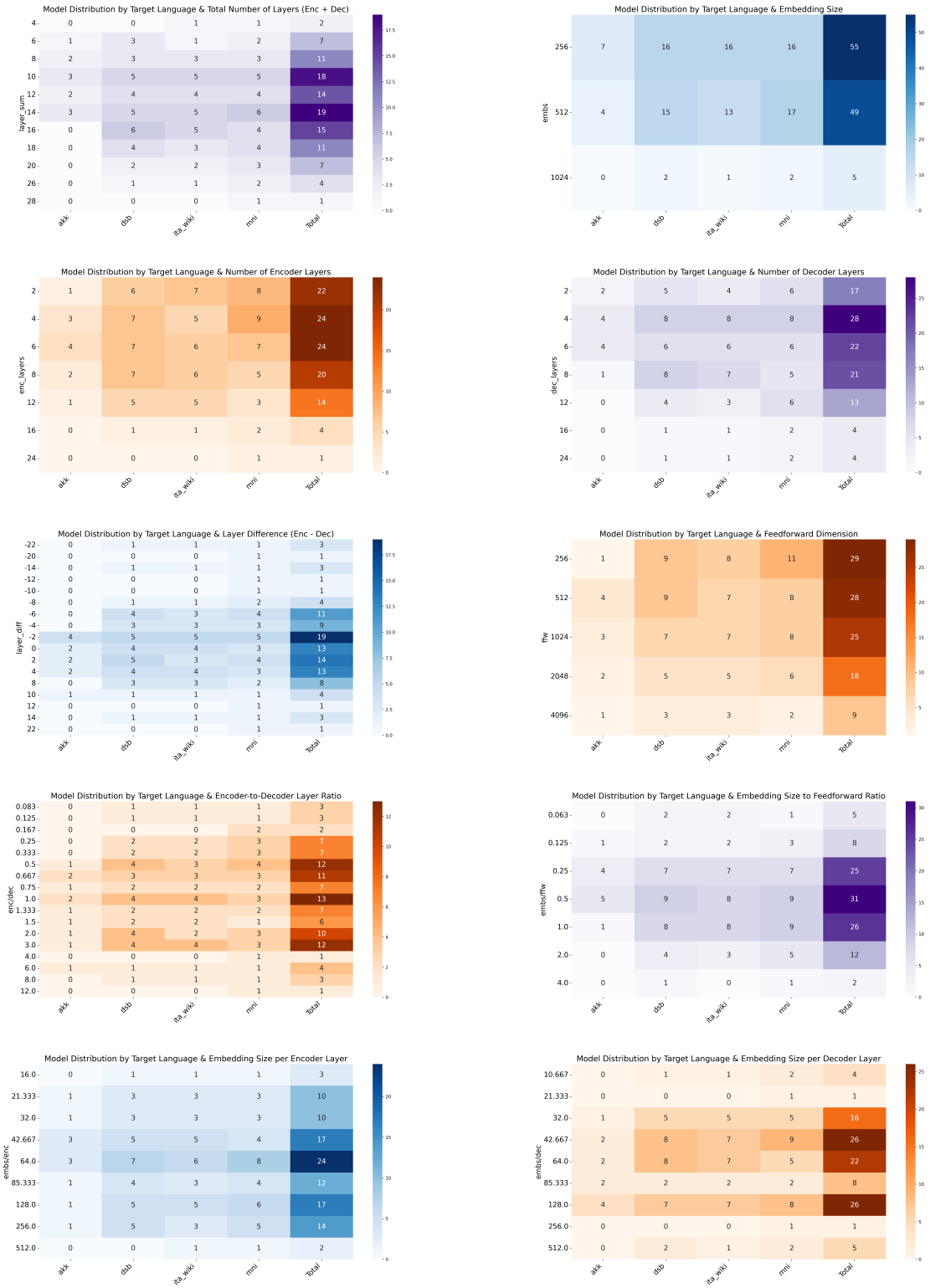

Figure 4: Heatmaps of the counts of configurations that achieve ChrF>35 with a PIES<1.

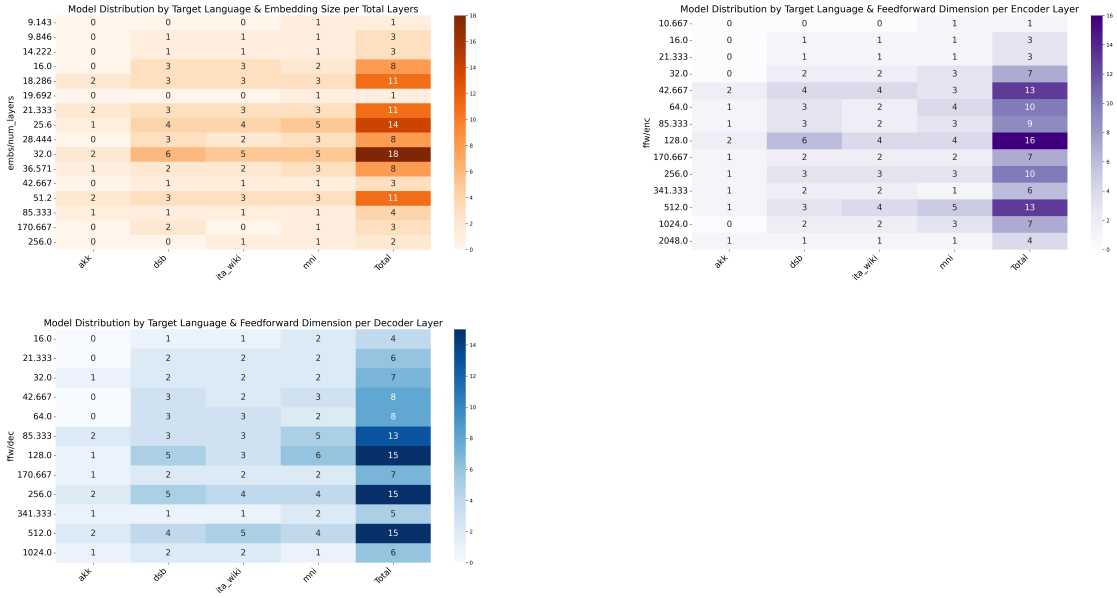

Figure 5: Heatmaps of the counts of configurations that achieve ChrF>35 with a PIES<1. (Continued)

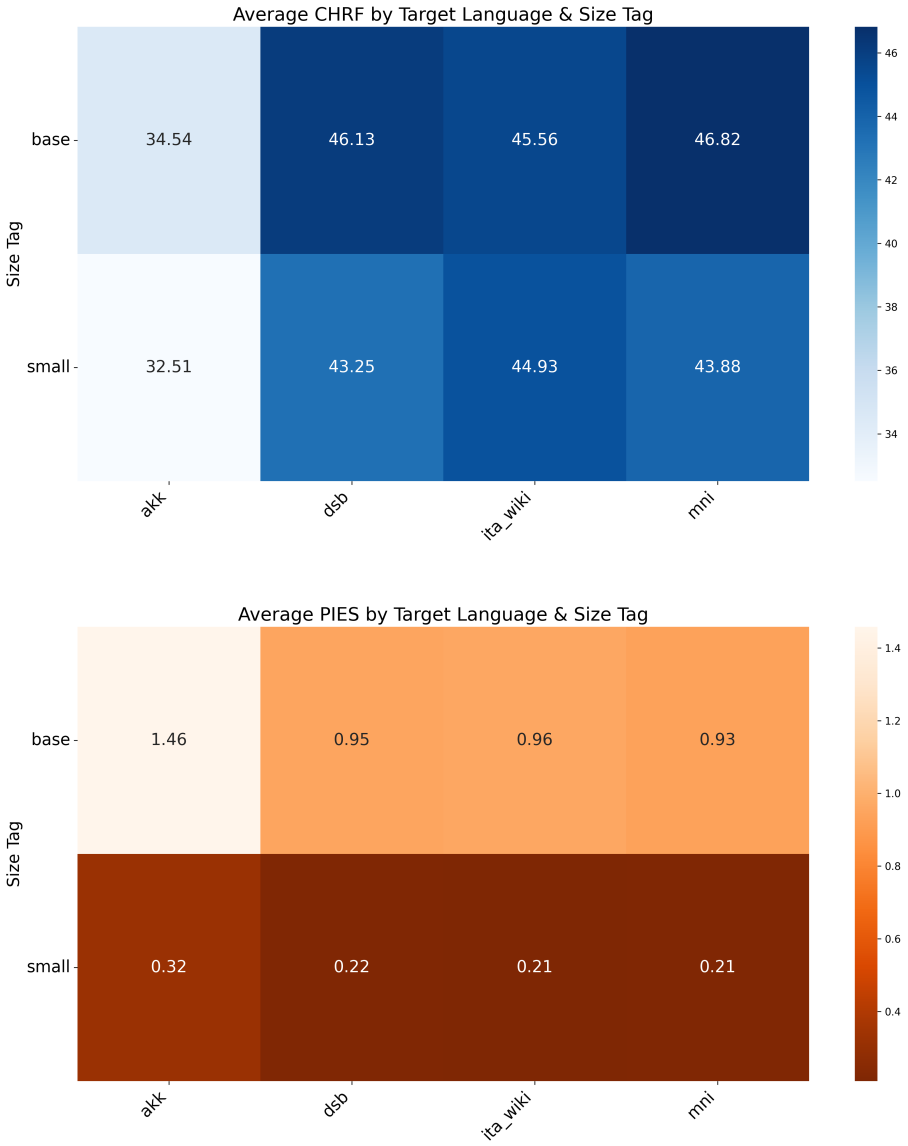

Figure 6: Average ChrF and PIES for the configurations in the optimal ranges, per size bracket. A lower PIES is better.

Table 12: Results for Experiment 1

src	tgt	model	bleu	chrf	comet	src	tgt	model	bleu	chrf	comet
eng	akk	2_4_256_1024_2	29.91	38.95	0.93	eng_wiki	ita_wiki	4_4_256_1024_2	12.59	43.86	0.59
eng	akk	4_2_256_1024_2	29.25	37.32	0.93	eng_wiki	ita_wiki	4_4_256_2048_2	13.38	45.08	0.61
eng	akk	4_4_256_256_2	30.23	39.68	0.93	eng_wiki	ita_wiki	4_4_256_4096_2	13.59	44.74	0.6
eng	akk	4_4_256_512_2	30.03	38.67	0.93	eng_wiki	ita_wiki	4_4_512_1024_2	13.53	45.28	0.6
eng	akk	4_4_256_1024_2	30.05	38.83	0.93	eng_wiki	ita_wiki	4_4_1024_1024_2	13.48	45.2	0.61
eng	akk	4_4_256_2048_2	29.82	38.43	0.93	eng_wiki	ita_wiki	4_4_2048_1024_2	14.04	45.61	0.62
eng	akk	4_4_256_4096_2	30.61	40.86	0.93	eng_wiki	ita_wiki	4_4_4096_1024_2	12.93	44.06	0.59
eng	akk	4_4_512_1024_2	30.33	40.49	0.93	eng_wiki	ita_wiki	4_6_256_1024_2	13.31	44.91	0.6
eng	akk	4_4_1024_1024_2	29.99	41.18	0.93	eng_wiki	ita_wiki	4_8_256_1024_2	13.57	44.86	0.6
eng	akk	4_4_2048_1024_2	30.17	41.79	0.93	eng_wiki	ita_wiki	4_12_256_1024_2	12.38	43.64	0.57
eng	akk	4_4_4096_1024_2	28.41	38.81	0.93	eng_wiki	ita_wiki	4_16_256_1024_2	13.07	44.33	0.58
eng	akk	4_6_256_1024_2	30.27	39.59	0.93	eng_wiki	ita_wiki	4_24_256_1024_2	13.57	45.23	0.61
eng	akk	4_8_256_1024_2	30.07	39.68	0.93	eng_wiki	ita_wiki	4_32_256_1024_2	2.18	19.42	0.26
eng	akk	4_12_256_1024_2	28.41	33.85	0.93	eng_wiki	ita_wiki	6_4_256_1024_2	13.28	44.79	0.61
eng	akk	4_16_256_1024_2	23.26	16.96	0.9	eng_wiki	ita_wiki	8_4_256_1024_2	13.74	45.15	0.61
eng	akk	4_24_256_1024_2	15.94	7.76	0.77	eng_wiki	ita_wiki	12_4_256_1024_2	13.55	44.86	0.61
eng	akk	4_32_256_1024_2	15.61	7.54	0.8	eng_wiki	ita_wiki	16_4_256_1024_2	3.68	24.42	0.36
eng	akk	6_4_256_1024_2	29.45	38.41	0.93	eng_wiki	ita_wiki	24_4_256_1024_2	3.26	21.55	0.35
eng	akk	8_4_256_1024_2	23.95	23.18	0.92	eng_wiki	ita_wiki	32_4_256_1024_2	2.81	19.07	0.34
eng	akk	12_4_256_1024_2	27.73	33.21	0.93	eng	mni	2_4_256_1024_2	16.15	44.33	0.69
eng	akk	16_4_256_1024_2	27.81	32.55	0.93	eng	mni	4_2_256_1024_2	16.93	44.01	0.69
eng	akk	24_4_256_1024_2	27.53	30.62	0.93	eng	mni	4_4_256_256_2	13.93	41.84	0.68
eng	akk	32_4_256_1024_2	21.49	16.32	0.89	eng	mni	4_4_256_512_2	16.06	43.39	0.69
deu	dsb	2_4_256_1024_2	28.06	43.7	0.63	eng	mni	4_4_256_1024_2	18.03	44.45	0.7
deu	dsb	4_2_256_1024_2	28.45	43.92	0.62	eng	mni	4_4_256_2048_2	18.2	44.75	0.7
deu	dsb	4_4_256_256_2	28.41	44.53	0.63	eng	mni	4_4_256_4096_2	20.45	46.11	0.7
deu	dsb	4_4_256_512_2	27.93	43.58	0.63	eng	mni	4_4_512_1024_2	20.2	47.35	0.71
deu	dsb	4_4_256_1024_2	28.09	43.01	0.62	eng	mni	4_4_1024_1024_2	21.88	48.5	0.71
deu	dsb	4_4_256_2048_2	29.28	45.19	0.64	eng	mni	4_4_2048_1024_2	21.45	48.36	0.7
deu	dsb	4_4_256_4096_2	28.53	43.95	0.63	eng	mni	4_4_4096_1024_2	18.7	43.87	0.68
deu	dsb	4_4_512_1024_2	29.34	46.72	0.64	eng	mni	4_6_256_1024_2	17.09	44.29	0.69
deu	dsb	4_4_1024_1024_2	29.96	48.29	0.65	eng	mni	4_8_256_1024_2	19.53	45.3	0.7
deu	dsb	4_4_2048_1024_2	30.48	48.88	0.65	eng	mni	4_12_256_1024_2	18.1	44.36	0.69
deu	dsb	4_4_4096_1024_2	29.8	48.28	0.64	eng	mni	4_16_256_1024_2	19.56	45.53	0.7
deu	dsb	4_6_256_1024_2	27.21	42.85	0.62	eng	mni	4_24_256_1024_2	18.84	45.54	0.69
deu	dsb	4_8_256_1024_2	27.65	42.74	0.62	eng	mni	4_32_256_1024_2	4.0	16.61	0.42
deu	dsb	4_12_256_1024_2	28.25	44.73	0.63	eng	mni	6_4_256_1024_2	18.39	44.71	0.7
deu	dsb	4_16_256_1024_2	27.34	44.02	0.63	eng	mni	8_4_256_1024_2	18.74	45.11	0.7
deu	dsb	4_24_256_1024_2	27.08	44.6	0.63	eng	mni	12_4_256_1024_2	15.59	39.42	0.68
deu	dsb	4_32_256_1024_2	11.53	25.88	0.51	eng	mni	16_4_256_1024_2	14.96	37.9	0.68
deu	dsb	6_4_256_1024_2	27.77	43.28	0.62	eng	mni	24_4_256_1024_2	13.63	35.27	0.66
deu	dsb	8_4_256_1024_2	27.82	43.06	0.62	eng	mni	32_4_256_1024_2	1.56	8.89	0.42
deu	dsb	12_4_256_1024_2	27.14	41.99	0.62						
deu	dsb	16_4_256_1024_2	16.14	22.42	0.52						
deu	dsb	24_4_256_1024_2	15.52	20.33	0.51						
deu	dsb	32_4_256_1024_2	13.81	15.04	0.48						
eng_wiki	ita_wiki	2_4_256_1024_2	12.77	44.22	0.59						
eng_wiki	ita_wiki	4_2_256_1024_2	13.72	45.05	0.62						
eng_wiki	ita_wiki	4_4_256_256_2	13.46	45.16	0.61						
eng_wiki	ita_wiki	4_4_256_512_2	13.65	45.14	0.61						

Table 13: Results for Experiment 2 (Part 1/6)

src	tgt	model	bleu	chrf	comet	src	tgt	model	bleu	chrf	comet
eng	akk	2_2_1024_1024_2	29.32	37.35	0.93	eng	akk	6_16_1024_512_2	29.16	38.74	0.93
eng	akk	2_2_2048_4096_2	29.53	39.33	0.93	eng	akk	6_24_256_2048_2	17.19	8.29	0.78
eng	akk	2_4_512_4096_2	31.08	42.57	0.93	eng	akk	6_24_512_4096_2	22.44	6.93	0.77
eng	akk	2_4_2048_256_2	29.88	40.84	0.93	eng	akk	6_32_256_1024_2	17.55	7.46	0.78
eng	akk	2_6_256_1024_2	26.61	28.13	0.93	eng	akk	8_4_256_512_2	29.67	38.25	0.93
eng	akk	2_8_256_512_2	25.76	26.4	0.92	eng	akk	8_4_512_2048_2	30.68	42.29	0.93
eng	akk	2_8_512_2048_2	30.31	41.23	0.94	eng	akk	8_4_1024_4096_2	30.18	42.55	0.93
eng	akk	2_8_1024_4096_2	29.94	41.77	0.93	eng	akk	8_6_256_256_2	23.57	14.41	0.88
eng	akk	2_12_512_1024_2	24.67	26.49	0.92	eng	akk	8_6_512_1024_2	30.06	40.84	0.93
eng	akk	2_12_1024_2048_2	30.48	42.58	0.93	eng	akk	8_8_256_4096_2	22.16	13.1	0.88
eng	akk	2_16_512_256_2	20.85	4.9	0.75	eng	akk	8_8_512_1024_2	21.25	15.6	0.88
eng	akk	2_16_1024_256_2	30.24	42.42	0.93	eng	akk	8_8_1024_2048_2	22.59	14.15	0.89
eng	akk	2_16_1024_512_2	30.1	41.81	0.93	eng	akk	8_12_512_256_2	20.22	14.54	0.89
eng	akk	2_16_1024_1024_2	30.81	43.07	0.93	eng	akk	8_12_512_512_2	20.67	15.01	0.9
eng	akk	2_24_256_2048_2	15.62	8.01	0.78	eng	akk	8_12_1024_512_2	21.77	14.94	0.88
eng	akk	2_24_512_4096_2	16.75	7.89	0.84	eng	akk	8_12_1024_1024_2	22.37	16.06	0.91
eng	akk	4_2_1024_256_2	30.08	40.15	0.93	eng	akk	8_16_1024_256_2	21.07	15.57	0.91
eng	akk	4_2_1024_512_2	29.27	37.93	0.93	eng	akk	8_32_256_1024_2	17.49	7.52	0.78
eng	akk	4_2_2048_512_2	28.78	37.81	0.93	eng	akk	8_32_512_2048_2	24.66	6.09	0.74
eng	akk	4_2_2048_1024_2	29.4	39.11	0.93	eng	akk	12_2_256_512_2	28.6	35.03	0.93
eng	akk	4_4_256_1024_2	30.18	38.92	0.93	eng	akk	12_2_1024_4096_2	28.09	37.46	0.93
eng	akk	4_6_256_512_2	29.87	38.81	0.93	eng	akk	12_4_256_256_2	26.68	28.97	0.93
eng	akk	4_6_512_2048_2	31.0	42.54	0.93	eng	akk	12_4_256_4096_2	28.29	34.0	0.93
eng	akk	4_6_1024_4096_2	29.65	42.02	0.93	eng	akk	12_4_512_1024_2	29.17	36.83	0.93
eng	akk	4_8_256_256_2	25.09	24.74	0.93	eng	akk	12_4_1024_2048_2	22.01	15.21	0.89
eng	akk	4_8_1024_4096_2	30.74	42.5	0.93	eng	akk	12_6_256_4096_2	27.56	30.33	0.93
eng	akk	4_12_256_4096_2	23.5	18.49	0.91	eng	akk	12_6_512_1024_2	29.52	37.17	0.93
eng	akk	4_12_512_512_2	22.87	13.21	0.89	eng	akk	12_6_1024_2048_2	29.04	38.79	0.93
eng	akk	4_12_1024_1024_2	30.45	42.45	0.93	eng	akk	12_8_512_512_2	28.55	35.78	0.93
eng	akk	4_12_1024_2048_2	30.45	42.74	0.93	eng	akk	12_8_1024_1024_2	20.41	16.14	0.9
eng	akk	4_16_512_256_2	30.3	39.35	0.93	eng	akk	12_12_512_256_2	22.66	18.15	0.91
eng	akk	4_16_1024_256_2	30.01	42.16	0.93	eng	akk	12_12_1024_256_2	21.42	18.64	0.91
eng	akk	4_16_1024_512_2	28.1	39.74	0.93	eng	akk	12_12_1024_512_2	22.83	16.31	0.9
eng	akk	4_24_256_2048_2	14.81	8.03	0.77	eng	akk	12_16_256_2048_2	17.54	8.26	0.78
eng	akk	4_24_512_4096_2	20.75	7.12	0.78	eng	akk	12_16_512_4096_2	19.53	10.25	0.84
eng	akk	6_2_256_1024_2	29.45	38.24	0.93	eng	akk	12_32_256_1024_2	17.82	9.23	0.8
eng	akk	6_2_512_4096_2	30.55	41.64	0.93	eng	akk	12_32_512_2048_2	18.07	9.72	0.82
eng	akk	6_2_2048_256_2	28.64	36.5	0.93	eng	akk	16_2_256_256_2	27.06	30.64	0.93
eng	akk	6_4_512_2048_2	29.93	39.88	0.94	eng	akk	16_2_256_4096_2	26.15	26.97	0.92
eng	akk	6_6_256_512_2	29.98	39.21	0.93	eng	akk	16_2_512_1024_2	27.54	34.01	0.93
eng	akk	6_6_512_2048_2	29.85	40.58	0.93	eng	akk	16_2_1024_2048_2	28.46	37.33	0.93
eng	akk	6_6_1024_4096_2	27.61	39.26	0.93	eng	akk	16_4_512_512_2	26.23	31.16	0.93
eng	akk	6_8_256_256_2	29.7	37.99	0.93	eng	akk	16_4_1024_2048_2	28.41	35.7	0.93
eng	akk	6_8_512_1024_2	30.04	40.92	0.93	eng	akk	16_6_512_512_2	28.35	35.79	0.93
eng	akk	6_8_1024_2048_2	30.97	43.39	0.93	eng	akk	16_6_1024_1024_2	26.32	28.94	0.91
eng	akk	6_12_256_4096_2	31.09	41.63	0.94	eng	akk	16_8_512_256_2	21.81	18.5	0.9
eng	akk	6_12_512_256_2	20.87	15.17	0.89	eng	akk	16_8_1024_512_2	20.68	17.16	0.89
eng	akk	6_12_512_512_2	22.29	21.44	0.92	eng	akk	16_8_1024_1024_2	28.35	36.26	0.92
eng	akk	6_12_1024_1024_2	22.22	14.15	0.9	eng	akk	16_12_256_2048_2	24.71	23.05	0.92
eng	akk	6_16_1024_256_2	30.27	41.0	0.94	eng	akk	16_12_512_4096_2	22.74	19.9	0.92

Table 14: Results for Experiment 2 (Part 2/6)

src	tgt	model	bleu	chrf	comet	src	tgt	model	bleu	chrf	comet
eng	akk	16_12_1024_256_2	20.4	15.52	0.89	deu	dsb	4_4_256_1024_2	28.09	43.01	0.62
eng	akk	16_16_256_2048_2	15.97	8.39	0.81	deu	dsb	4_6_256_512_2	26.91	42.86	0.62
eng	akk	16_16_512_4096_2	16.2	11.54	0.81	deu	dsb	4_6_512_2048_2	28.82	45.93	0.63
eng	akk	16_24_256_1024_2	17.39	8.28	0.8	deu	dsb	4_6_1024_4096_2	31.57	49.65	0.65
eng	akk	16_32_256_1024_2	24.86	6.13	0.74	deu	dsb	4_8_256_256_2	28.06	43.14	0.62
eng	akk	16_32_512_2048_2	16.7	7.76	0.82	deu	dsb	4_8_1024_4096_2	32.1	50.21	0.66
eng	akk	24_2_512_256_2	27.45	32.36	0.93	deu	dsb	4_12_256_4096_2	27.49	42.94	0.62
eng	akk	24_2_512_512_2	27.56	33.44	0.93	deu	dsb	4_12_512_512_2	28.78	45.54	0.63
eng	akk	24_2_1024_1024_2	27.51	33.86	0.93	deu	dsb	4_12_1024_1024_2	30.11	47.66	0.65
eng	akk	24_4_256_2048_2	21.96	12.81	0.9	deu	dsb	4_12_1024_2048_2	31.2	48.69	0.65
eng	akk	24_4_512_256_2	25.96	31.73	0.93	deu	dsb	4_16_512_256_2	28.8	46.63	0.64
eng	akk	24_4_1024_512_2	27.35	33.82	0.92	deu	dsb	4_16_1024_256_2	29.46	47.91	0.65
eng	akk	24_6_256_2048_2	21.88	11.6	0.87	deu	dsb	4_16_1024_512_2	30.57	48.87	0.65
eng	akk	24_6_512_4096_2	19.84	8.84	0.85	deu	dsb	4_24_256_2048_2	12.22	24.03	0.5
eng	akk	24_6_1024_256_2	27.85	35.54	0.92	deu	dsb	4_24_512_4096_2	15.34	29.54	0.53
eng	akk	24_6_1024_512_2	23.13	22.75	0.9	deu	dsb	6_2_256_1024_2	27.86	44.0	0.63
eng	akk	24_8_256_2048_2	21.87	8.91	0.82	deu	dsb	6_2_512_4096_2	29.87	47.84	0.64
eng	akk	24_8_512_4096_2	20.25	9.48	0.86	deu	dsb	6_2_2048_256_2	28.13	46.89	0.64
eng	akk	24_8_1024_256_2	26.93	32.41	0.92	deu	dsb	6_4_512_2048_2	28.34	45.13	0.64
eng	akk	24_24_256_1024_2	17.66	8.16	0.82	deu	dsb	6_6_256_512_2	27.4	42.87	0.62
eng	akk	24_24_512_2048_2	24.81	6.16	0.74	deu	dsb	6_6_512_2048_2	30.05	46.83	0.64
eng	akk	24_32_256_512_2	19.73	8.9	0.8	deu	dsb	6_6_1024_4096_2	32.08	50.72	0.66
eng	akk	32_2_256_2048_2	21.87	8.76	0.83	deu	dsb	6_8_256_256_2	27.48	42.48	0.62
eng	akk	32_2_512_4096_2	22.38	7.18	0.86	deu	dsb	6_8_512_1024_2	28.41	45.33	0.63
eng	akk	32_2_1024_256_2	20.99	15.41	0.9	deu	dsb	6_8_1024_2048_2	33.03	50.91	0.66
eng	akk	32_12_256_1024_2	23.7	17.46	0.91	deu	dsb	6_12_256_4096_2	28.67	44.24	0.63
eng	akk	32_16_256_1024_2	14.03	7.99	0.77	deu	dsb	6_12_512_256_2	29.46	46.16	0.64
eng	akk	32_16_512_2048_2	18.57	8.35	0.83	deu	dsb	6_12_512_512_2	28.63	45.56	0.64
eng	akk	32_32_256_512_2	19.11	8.39	0.82	deu	dsb	6_12_1024_1024_2	31.58	49.56	0.65
eng	akk	32_32_256_4096_2	24.52	6.24	0.74	deu	dsb	6_16_1024_256_2	30.51	48.94	0.65
eng	akk	32_32_512_1024_2	16.92	8.25	0.82	deu	dsb	6_16_1024_512_2	31.32	50.22	0.66
deu	dsb	2_2_2048_4096_2	28.96	46.8	0.65	deu	dsb	6_24_256_2048_2	10.04	13.03	0.46
deu	dsb	2_4_512_4096_2	29.43	46.96	0.64	deu	dsb	6_24_512_4096_2	8.71	15.32	0.45
deu	dsb	2_4_2048_256_2	29.03	47.44	0.64	deu	dsb	6_32_256_1024_2	10.96	24.02	0.5
deu	dsb	2_6_256_1024_2	28.35	44.33	0.63	deu	dsb	8_4_256_512_2	26.89	42.42	0.62
deu	dsb	2_8_256_512_2	28.84	43.92	0.63	deu	dsb	8_4_512_2048_2	29.05	46.24	0.64
deu	dsb	2_8_512_2048_2	27.02	44.16	0.63	deu	dsb	8_4_1024_4096_2	32.82	51.13	0.67
deu	dsb	2_8_1024_4096_2	29.98	47.89	0.65	deu	dsb	8_6_256_256_2	28.06	43.52	0.62
deu	dsb	2_12_512_1024_2	28.46	45.47	0.63	deu	dsb	8_6_512_1024_2	30.38	46.91	0.64
deu	dsb	2_12_1024_2048_2	29.99	47.56	0.64	deu	dsb	8_8_256_4096_2	27.89	43.46	0.62
deu	dsb	2_16_512_256_2	29.11	46.67	0.64	deu	dsb	8_8_512_1024_2	31.31	47.63	0.65
deu	dsb	2_16_1024_256_2	29.41	47.41	0.64	deu	dsb	8_8_1024_2048_2	31.24	49.36	0.65
deu	dsb	2_16_1024_512_2	30.36	48.81	0.65	deu	dsb	8_12_512_256_2	28.07	45.09	0.63
deu	dsb	2_16_1024_1024_2	30.68	49.03	0.65	deu	dsb	8_12_512_512_2	28.69	45.59	0.64
deu	dsb	2_24_256_2048_2	27.0	44.05	0.63	deu	dsb	8_12_1024_512_2	31.46	49.17	0.66
deu	dsb	2_24_512_4096_2	29.72	47.3	0.65	deu	dsb	8_12_1024_1024_2	31.66	50.01	0.66
deu	dsb	4_2_1024_256_2	29.47	46.86	0.64	deu	dsb	8_16_1024_256_2	31.34	50.58	0.66
deu	dsb	4_2_1024_512_2	29.71	48.32	0.65	deu	dsb	8_32_256_1024_2	11.97	26.74	0.51
deu	dsb	4_2_2048_512_2	27.07	45.25	0.63	deu	dsb	8_32_512_2048_2	13.85	30.58	0.53
deu	dsb	4_2_2048_1024_2	28.91	47.68	0.64	deu	dsb	12_2_256_512_2	27.53	42.04	0.61

Table 15: Results for Experiment 2 (Part 3/6)

src	tgt	model	bleu	chrf	comet	src	tgt	model	bleu	chrf	comet
deu	dsb	12_2_1024_4096_2	33.18	51.57	0.67	deu	dsb	24_24_512_2048_2	9.03	3.64	0.4
deu	dsb	12_4_256_256_2	28.59	43.07	0.62	deu	dsb	24_32_256_512_2	5.29	3.32	0.38
deu	dsb	12_4_256_4096_2	26.51	41.66	0.62	deu	dsb	32_2_256_2048_2	11.11	6.02	0.44
deu	dsb	12_4_512_1024_2	29.51	46.58	0.64	deu	dsb	32_2_512_4096_2	5.8	5.57	0.42
deu	dsb	12_4_1024_2048_2	30.87	49.04	0.65	deu	dsb	32_2_1024_256_2	15.96	22.55	0.52
deu	dsb	12_6_256_4096_2	26.13	40.19	0.6	deu	dsb	32_12_256_1024_2	12.44	10.4	0.46
deu	dsb	12_6_512_1024_2	30.35	47.13	0.65	deu	dsb	32_16_256_1024_2	12.44	10.01	0.46
deu	dsb	12_6_1024_2048_2	30.34	47.96	0.65	deu	dsb	32_16_512_2048_2	10.36	9.99	0.44
deu	dsb	12_8_512_512_2	28.85	44.68	0.63	deu	dsb	32_32_256_512_2	10.26	3.02	0.39
deu	dsb	12_8_1024_1024_2	30.41	48.14	0.65	deu	dsb	32_32_256_4096_2	5.24	3.13	0.41
deu	dsb	12_12_512_256_2	28.8	45.28	0.64	deu	dsb	32_32_512_1024_2	10.79	2.9	0.4
deu	dsb	12_12_1024_256_2	31.39	48.98	0.65	eng_wiki	ita_wiki	2_2_1024_1024_2	12.56	44.02	0.59
deu	dsb	12_12_1024_512_2	32.07	49.79	0.66	eng_wiki	ita_wiki	2_2_2048_4096_2	13.56	44.71	0.61
deu	dsb	12_16_256_2048_2	26.18	40.04	0.61	eng_wiki	ita_wiki	2_4_512_4096_2	13.86	44.95	0.61
deu	dsb	12_16_512_4096_2	26.25	41.22	0.61	eng_wiki	ita_wiki	2_4_2048_256_2	12.37	43.23	0.57
deu	dsb	12_32_256_1024_2	10.84	19.88	0.49	eng_wiki	ita_wiki	2_6_256_1024_2	13.43	44.79	0.61
deu	dsb	12_32_512_2048_2	7.47	3.75	0.4	eng_wiki	ita_wiki	2_8_256_512_2	13.1	44.67	0.6
deu	dsb	16_2_256_256_2	24.96	37.98	0.6	eng_wiki	ita_wiki	2_8_512_2048_2	12.93	44.14	0.59
deu	dsb	16_2_256_4096_2	15.16	17.72	0.49	eng_wiki	ita_wiki	2_8_1024_4096_2	13.74	44.96	0.61
deu	dsb	16_2_512_1024_2	17.76	24.74	0.54	eng_wiki	ita_wiki	2_12_512_1024_2	12.54	43.97	0.59
deu	dsb	16_2_1024_2048_2	18.64	28.2	0.55	eng_wiki	ita_wiki	2_12_1024_2048_2	13.1	44.71	0.61
deu	dsb	16_4_512_512_2	17.65	25.5	0.53	eng_wiki	ita_wiki	2_16_512_256_2	13.04	45.0	0.59
deu	dsb	16_4_1024_2048_2	17.98	27.12	0.55	eng_wiki	ita_wiki	2_16_1024_256_2	13.42	45.15	0.61
deu	dsb	16_6_512_512_2	18.54	26.36	0.54	eng_wiki	ita_wiki	2_16_1024_512_2	13.19	44.68	0.61
deu	dsb	16_6_1024_1024_2	17.22	27.33	0.54	eng_wiki	ita_wiki	2_16_1024_1024_2	13.42	44.85	0.61
deu	dsb	16_8_512_256_2	18.76	26.24	0.54	eng_wiki	ita_wiki	2_24_256_2048_2	12.31	43.76	0.58
deu	dsb	16_8_1024_512_2	18.9	28.61	0.55	eng_wiki	ita_wiki	2_24_512_4096_2	3.16	24.5	0.28
deu	dsb	16_8_1024_1024_2	18.49	28.1	0.55	eng_wiki	ita_wiki	4_2_1024_256_2	12.73	44.67	0.59
deu	dsb	16_12_256_2048_2	16.13	21.51	0.51	eng_wiki	ita_wiki	4_2_1024_512_2	12.76	44.19	0.59
deu	dsb	16_12_512_4096_2	17.47	25.2	0.53	eng_wiki	ita_wiki	4_2_2048_512_2	13.04	44.83	0.59
deu	dsb	16_12_1024_256_2	18.15	29.32	0.55	eng_wiki	ita_wiki	4_2_2048_1024_2	13.91	45.42	0.62
deu	dsb	16_16_256_2048_2	16.62	20.96	0.51	eng_wiki	ita_wiki	4_4_256_1024_2	13.4	44.92	0.61
deu	dsb	16_16_512_4096_2	17.78	24.4	0.53	eng_wiki	ita_wiki	4_6_256_512_2	13.68	45.07	0.61
deu	dsb	16_24_256_1024_2	8.87	3.16	0.4	eng_wiki	ita_wiki	4_6_512_2048_2	14.15	45.72	0.62
deu	dsb	16_32_256_1024_2	6.44	4.5	0.39	eng_wiki	ita_wiki	4_6_1024_4096_2	14.7	46.49	0.65
deu	dsb	16_32_512_2048_2	10.55	4.15	0.41	eng_wiki	ita_wiki	4_8_256_256_2	12.97	44.34	0.59
deu	dsb	24_2_512_256_2	16.25	24.86	0.52	eng_wiki	ita_wiki	4_8_1024_4096_2	14.53	45.99	0.63
deu	dsb	24_2_512_512_2	16.83	24.89	0.52	eng_wiki	ita_wiki	4_12_256_4096_2	11.88	42.9	0.56
deu	dsb	24_2_1024_1024_2	18.11	25.05	0.54	eng_wiki	ita_wiki	4_12_512_512_2	13.4	45.14	0.6
deu	dsb	24_4_256_2048_2	12.05	10.72	0.46	eng_wiki	ita_wiki	4_12_1024_1024_2	13.84	45.25	0.62
deu	dsb	24_4_512_256_2	17.07	24.17	0.53	eng_wiki	ita_wiki	4_12_1024_2048_2	13.91	45.54	0.62
deu	dsb	24_4_1024_512_2	17.84	26.84	0.54	eng_wiki	ita_wiki	4_16_512_256_2	13.58	45.21	0.62
deu	dsb	24_6_256_2048_2	13.07	12.1	0.47	eng_wiki	ita_wiki	4_16_1024_256_2	13.92	45.8	0.62
deu	dsb	24_6_512_4096_2	11.96	15.7	0.47	eng_wiki	ita_wiki	4_16_1024_512_2	14.01	45.91	0.62
deu	dsb	24_6_1024_256_2	18.5	24.87	0.54	eng_wiki	ita_wiki	4_24_256_2048_2	3.22	26.13	0.28
deu	dsb	24_6_1024_512_2	18.55	26.37	0.55	eng_wiki	ita_wiki	4_24_512_4096_2	2.35	22.59	0.26
deu	dsb	24_8_256_2048_2	12.73	12.56	0.46	eng_wiki	ita_wiki	6_2_256_1024_2	13.81	45.35	0.61
deu	dsb	24_8_512_4096_2	13.9	16.93	0.5	eng_wiki	ita_wiki	6_2_512_4096_2	14.02	45.48	0.62
deu	dsb	24_8_1024_256_2	17.26	26.91	0.53	eng_wiki	ita_wiki	6_2_2048_256_2	14.07	45.9	0.62
deu	dsb	24_24_256_1024_2	7.41	4.06	0.41	eng_wiki	ita_wiki	6_4_512_2048_2	14.59	46.28	0.63

Table 16: Results for Experiment 2 (Part 4/6)

src	tgt	model	bleu	chrf	comet	src	tgt	model	bleu	chrf	comet
eng_wiki	ita_wiki	6_6_256_512_2	13.51	45.14	0.61	eng_wiki	ita_wiki	16_2_512_1024_2	3.76	25.12	0.35
eng_wiki	ita_wiki	6_6_512_2048_2	14.27	45.57	0.61	eng_wiki	ita_wiki	16_2_1024_2048_2	4.01	26.05	0.35
eng_wiki	ita_wiki	6_6_1024_4096_2	15.27	46.69	0.65	eng_wiki	ita_wiki	16_4_512_512_2	4.03	26.47	0.37
eng_wiki	ita_wiki	6_8_256_256_2	13.61	45.16	0.61	eng_wiki	ita_wiki	16_4_1024_2048_2	3.89	26.76	0.38
eng_wiki	ita_wiki	6_8_512_1024_2	13.73	45.33	0.61	eng_wiki	ita_wiki	16_6_512_512_2	4.05	25.89	0.38
eng_wiki	ita_wiki	6_8_1024_2048_2	14.4	46.14	0.65	eng_wiki	ita_wiki	16_6_1024_1024_2	4.02	27.6	0.38
eng_wiki	ita_wiki	6_12_256_4096_2	12.2	43.39	0.56	eng_wiki	ita_wiki	16_8_512_256_2	3.96	27.52	0.38
eng_wiki	ita_wiki	6_12_512_256_2	13.54	45.62	0.61	eng_wiki	ita_wiki	16_8_1024_512_2	4.22	28.34	0.38
eng_wiki	ita_wiki	6_12_512_512_2	13.39	45.05	0.61	eng_wiki	ita_wiki	16_8_1024_1024_2	4.0	28.29	0.38
eng_wiki	ita_wiki	6_12_1024_1024_2	14.61	46.4	0.64	eng_wiki	ita_wiki	16_12_256_2048_2	3.48	23.67	0.37
eng_wiki	ita_wiki	6_16_1024_256_2	14.4	46.27	0.63	eng_wiki	ita_wiki	16_12_512_4096_2	3.63	25.0	0.38
eng_wiki	ita_wiki	6_16_1024_512_2	14.41	46.48	0.64	eng_wiki	ita_wiki	16_12_1024_256_2	4.16	28.04	0.39
eng_wiki	ita_wiki	6_24_256_2048_2	3.03	25.75	0.28	eng_wiki	ita_wiki	16_16_256_2048_2	3.47	23.38	0.37
eng_wiki	ita_wiki	6_24_512_4096_2	2.25	20.31	0.25	eng_wiki	ita_wiki	16_16_512_4096_2	3.69	24.9	0.38
eng_wiki	ita_wiki	6_32_256_1024_2	2.59	23.99	0.27	eng_wiki	ita_wiki	16_24_256_1024_2	1.41	5.47	0.22
eng_wiki	ita_wiki	8_4_256_512_2	13.44	45.07	0.6	eng_wiki	ita_wiki	16_32_256_1024_2	1.33	2.99	0.21
eng_wiki	ita_wiki	8_4_512_2048_2	14.38	46.27	0.63	eng_wiki	ita_wiki	16_32_512_2048_2	1.38	6.65	0.22
eng_wiki	ita_wiki	8_4_1024_4096_2	15.74	47.32	0.67	eng_wiki	ita_wiki	24_2_512_256_2	3.79	24.36	0.35
eng_wiki	ita_wiki	8_6_256_256_2	13.37	44.82	0.6	eng_wiki	ita_wiki	24_2_512_512_2	3.65	23.5	0.34
eng_wiki	ita_wiki	8_6_512_1024_2	13.99	45.65	0.62	eng_wiki	ita_wiki	24_2_1024_1024_2	3.69	25.0	0.37
eng_wiki	ita_wiki	8_8_256_4096_2	12.89	44.41	0.58	eng_wiki	ita_wiki	24_4_256_2048_2	3.04	21.33	0.36
eng_wiki	ita_wiki	8_8_512_1024_2	13.97	45.77	0.62	eng_wiki	ita_wiki	24_4_512_256_2	3.64	25.14	0.37
eng_wiki	ita_wiki	8_8_1024_2048_2	15.01	46.59	0.65	eng_wiki	ita_wiki	24_4_1024_512_2	3.73	24.99	0.36
eng_wiki	ita_wiki	8_12_512_256_2	13.7	45.62	0.61	eng_wiki	ita_wiki	24_6_256_2048_2	2.86	20.14	0.35
eng_wiki	ita_wiki	8_12_512_512_2	13.59	45.17	0.61	eng_wiki	ita_wiki	24_6_512_4096_2	2.85	20.86	0.38
eng_wiki	ita_wiki	8_12_1024_512_2	14.92	46.66	0.65	eng_wiki	ita_wiki	24_6_1024_256_2	3.77	26.25	0.38
eng_wiki	ita_wiki	8_12_1024_1024_2	15.06	46.97	0.65	eng_wiki	ita_wiki	24_6_1024_512_2	3.59	26.53	0.38
eng_wiki	ita_wiki	8_16_1024_256_2	14.88	47.08	0.65	eng_wiki	ita_wiki	24_8_256_2048_2	2.93	20.88	0.36
eng_wiki	ita_wiki	8_32_256_1024_2	3.23	26.48	0.28	eng_wiki	ita_wiki	24_8_512_4096_2	2.93	22.32	0.38
eng_wiki	ita_wiki	8_32_512_2048_2	3.68	28.03	0.29	eng_wiki	ita_wiki	24_8_1024_256_2	3.57	26.02	0.38
eng_wiki	ita_wiki	12_2_256_512_2	13.02	44.15	0.59	eng_wiki	ita_wiki	24_24_256_1024_2	1.81	2.93	0.22
eng_wiki	ita_wiki	12_2_1024_4096_2	16.12	47.89	0.66	eng_wiki	ita_wiki	24_24_512_2048_2	1.38	2.46	0.22
eng_wiki	ita_wiki	12_4_256_256_2	13.52	45.23	0.61	eng_wiki	ita_wiki	24_32_256_512_2	1.48	7.86	0.23
eng_wiki	ita_wiki	12_4_256_4096_2	13.37	44.22	0.59	eng_wiki	ita_wiki	32_2_256_2048_2	2.56	18.81	0.33
eng_wiki	ita_wiki	12_4_512_1024_2	13.87	45.61	0.62	eng_wiki	ita_wiki	32_2_512_4096_2	2.24	17.43	0.33
eng_wiki	ita_wiki	12_4_1024_2048_2	15.61	47.63	0.66	eng_wiki	ita_wiki	32_2_1024_256_2	2.76	18.58	0.34
eng_wiki	ita_wiki	12_6_256_4096_2	12.64	43.68	0.58	eng_wiki	ita_wiki	32_12_256_1024_2	2.68	20.44	0.34
eng_wiki	ita_wiki	12_6_512_1024_2	14.0	45.74	0.61	eng_wiki	ita_wiki	32_16_256_1024_2	2.67	20.64	0.35
eng_wiki	ita_wiki	12_6_1024_2048_2	15.34	47.16	0.66	eng_wiki	ita_wiki	32_16_512_2048_2	2.46	19.93	0.36
eng_wiki	ita_wiki	12_8_512_512_2	14.28	46.03	0.62	eng_wiki	ita_wiki	32_32_256_512_2	1.66	3.55	0.22
eng_wiki	ita_wiki	12_8_1024_1024_2	14.57	46.41	0.63	eng_wiki	ita_wiki	32_32_256_4096_2	1.41	1.6	0.2
eng_wiki	ita_wiki	12_12_512_256_2	13.37	45.32	0.6	eng_wiki	ita_wiki	32_32_512_1024_2	1.55	2.37	0.21
eng_wiki	ita_wiki	12_12_1024_256_2	15.24	47.12	0.64	eng	mni	2_2_1024_1024_2	20.5	47.35	0.7
eng_wiki	ita_wiki	12_12_1024_512_2	14.67	46.45	0.64	eng	mni	2_2_2048_4096_2	21.32	47.97	0.7
eng_wiki	ita_wiki	12_16_256_2048_2	11.27	42.02	0.54	eng	mni	2_4_512_4096_2	20.97	47.45	0.7
eng_wiki	ita_wiki	12_16_512_4096_2	11.97	42.82	0.56	eng	mni	2_4_2048_256_2	21.57	48.18	0.7
eng_wiki	ita_wiki	12_32_256_1024_2	1.29	3.86	0.21	eng	mni	2_6_256_1024_2	19.46	45.8	0.7
eng_wiki	ita_wiki	12_32_512_2048_2	3.77	27.69	0.29	eng	mni	2_8_256_512_2	18.62	45.96	0.7
eng_wiki	ita_wiki	16_2_256_256_2	11.92	42.5	0.56	eng	mni	2_8_512_2048_2	20.65	47.48	0.7
eng_wiki	ita_wiki	16_2_256_4096_2	3.27	21.27	0.35	eng	mni	2_8_1024_4096_2	21.7	48.53	0.71

Table 17: Results for Experiment 2 (Part 5/6)

src	tgt	model	bleu	chrf	comet	src	tgt	model	bleu	chrf	comet
eng	mni	2_12_512_1024_2	20.96	48.15	0.71	eng	mni	8_6_512_1024_2	20.32	47.29	0.71
eng	mni	2_12_1024_2048_2	21.94	49.38	0.71	eng	mni	8_8_256_4096_2	18.93	41.98	0.68
eng	mni	2_16_512_256_2	20.96	47.99	0.71	eng	mni	8_8_512_1024_2	19.81	46.04	0.7
eng	mni	2_16_1024_256_2	21.51	49.88	0.71	eng	mni	8_8_1024_2048_2	21.33	48.04	0.71
eng	mni	2_16_1024_512_2	21.52	49.7	0.71	eng	mni	8_12_512_256_2	19.55	46.84	0.71
eng	mni	2_16_1024_1024_2	21.68	48.97	0.71	eng	mni	8_12_512_512_2	20.57	47.32	0.71
eng	mni	2_24_256_2048_2	18.53	45.07	0.69	eng	mni	8_12_1024_512_2	21.6	48.25	0.71
eng	mni	2_24_512_4096_2	21.14	48.14	0.7	eng	mni	8_12_1024_1024_2	21.16	47.83	0.71
eng	mni	4_2_1024_256_2	20.08	46.9	0.7	eng	mni	8_16_1024_256_2	20.48	47.36	0.71
eng	mni	4_2_1024_512_2	20.46	47.53	0.7	eng	mni	8_32_256_1024_2	4.71	18.75	0.45
eng	mni	4_2_2048_512_2	19.91	46.03	0.69	eng	mni	8_32_512_2048_2	8.73	29.97	0.54
eng	mni	4_2_2048_1024_2	20.23	46.64	0.69	eng	mni	12_2_256_512_2	14.79	39.85	0.68
eng	mni	4_4_256_1024_2	17.38	44.5	0.69	eng	mni	12_2_1024_4096_2	19.94	45.84	0.7
eng	mni	4_6_256_512_2	14.69	41.51	0.68	eng	mni	12_4_256_256_2	15.01	39.36	0.68
eng	mni	4_6_512_2048_2	20.61	47.51	0.7	eng	mni	12_4_256_4096_2	17.39	38.81	0.68
eng	mni	4_6_1024_4096_2	22.28	49.07	0.71	eng	mni	12_4_512_1024_2	19.2	44.07	0.7
eng	mni	4_8_256_256_2	17.69	44.48	0.69	eng	mni	12_4_1024_2048_2	19.98	44.33	0.69
eng	mni	4_8_1024_4096_2	22.02	49.02	0.71	eng	mni	12_6_256_4096_2	17.57	38.62	0.68
eng	mni	4_12_256_4096_2	19.78	45.44	0.7	eng	mni	12_6_512_1024_2	19.01	43.38	0.69
eng	mni	4_12_512_512_2	20.77	47.58	0.71	eng	mni	12_6_1024_2048_2	19.39	44.1	0.69
eng	mni	4_12_1024_1024_2	21.61	49.21	0.71	eng	mni	12_8_512_512_2	18.57	43.11	0.69
eng	mni	4_12_1024_2048_2	21.86	48.87	0.71	eng	mni	12_8_1024_1024_2	19.74	44.36	0.69
eng	mni	4_16_512_256_2	20.2	47.73	0.71	eng	mni	12_12_512_256_2	18.94	43.92	0.7
eng	mni	4_16_1024_256_2	21.86	48.67	0.71	eng	mni	12_12_1024_256_2	20.07	44.85	0.7
eng	mni	4_16_1024_512_2	21.74	49.41	0.71	eng	mni	12_12_1024_512_2	19.2	45.0	0.7
eng	mni	4_24_256_2048_2	19.64	45.94	0.7	eng	mni	12_16_256_2048_2	17.19	39.29	0.68
eng	mni	4_24_512_4096_2	4.15	16.81	0.42	eng	mni	12_16_512_4096_2	19.34	43.19	0.69
eng	mni	6_2_256_1024_2	16.71	43.44	0.69	eng	mni	12_32_256_1024_2	3.96	16.01	0.42
eng	mni	6_2_512_4096_2	20.56	47.24	0.7	eng	mni	12_32_512_2048_2	3.51	16.37	0.41
eng	mni	6_2_2048_256_2	19.46	45.47	0.69	eng	mni	16_2_256_256_2	12.51	36.0	0.66
eng	mni	6_4_512_2048_2	20.61	47.66	0.71	eng	mni	16_2_256_4096_2	15.8	36.51	0.66
eng	mni	6_6_256_512_2	17.38	44.22	0.69	eng	mni	16_2_512_1024_2	17.12	41.31	0.69
eng	mni	6_6_512_2048_2	20.14	46.7	0.7	eng	mni	16_2_1024_2048_2	19.83	43.59	0.69
eng	mni	6_6_1024_4096_2	21.54	48.65	0.71	eng	mni	16_4_512_512_2	18.0	41.83	0.69
eng	mni	6_8_256_256_2	17.31	44.25	0.7	eng	mni	16_4_1024_2048_2	18.93	43.39	0.68
eng	mni	6_8_512_1024_2	20.77	48.13	0.71	eng	mni	16_6_512_512_2	17.97	42.24	0.69
eng	mni	6_8_1024_2048_2	21.45	47.59	0.7	eng	mni	16_6_1024_1024_2	18.44	42.31	0.68
eng	mni	6_12_256_4096_2	19.68	45.01	0.7	eng	mni	16_8_512_256_2	17.6	41.4	0.69
eng	mni	6_12_512_256_2	20.82	48.02	0.71	eng	mni	16_8_1024_512_2	19.05	42.87	0.68
eng	mni	6_12_512_512_2	20.82	48.3	0.71	eng	mni	16_8_1024_1024_2	19.36	42.82	0.68
eng	mni	6_12_1024_1024_2	21.23	49.09	0.71	eng	mni	16_12_256_2048_2	16.53	38.3	0.67
eng	mni	6_16_1024_256_2	21.32	48.92	0.71	eng	mni	16_12_512_4096_2	18.58	41.19	0.68
eng	mni	6_16_1024_512_2	21.22	48.73	0.71	eng	mni	16_12_1024_256_2	19.05	43.73	0.69
eng	mni	6_24_256_2048_2	19.62	46.48	0.7	eng	mni	16_16_256_2048_2	17.28	38.61	0.67
eng	mni	6_24_512_4096_2	4.63	17.12	0.43	eng	mni	16_16_512_4096_2	19.24	41.86	0.68
eng	mni	6_32_256_1024_2	2.69	11.42	0.37	eng	mni	16_24_256_1024_2	2.55	7.92	0.39
eng	mni	8_4_256_512_2	15.21	41.64	0.68	eng	mni	16_32_256_1024_2	3.52	14.28	0.4
eng	mni	8_4_512_2048_2	20.79	47.27	0.71	eng	mni	16_32_512_2048_2	2.8	12.5	0.4
eng	mni	8_4_1024_4096_2	21.31	47.73	0.7	eng	mni	24_2_512_256_2	15.61	40.47	0.68
eng	mni	8_6_256_256_2	16.37	42.96	0.69	eng	mni	24_2_512_512_2	16.97	40.52	0.68

Table 18: Results for Experiment 2 (Part 6/6)

src	tgt	model	bleu	chrf	comet
eng	mni	24_2_1024_1024_2	18.43	42.35	0.68
eng	mni	24_4_256_2048_2	14.6	34.64	0.66
eng	mni	24_4_512_256_2	15.55	38.65	0.67
eng	mni	24_4_1024_512_2	18.81	42.09	0.68
eng	mni	24_6_256_2048_2	15.31	34.31	0.66
eng	mni	24_6_512_4096_2	15.34	33.19	0.65
eng	mni	24_6_1024_256_2	18.17	41.86	0.68
eng	mni	24_6_1024_512_2	18.12	41.28	0.68
eng	mni	24_8_256_2048_2	15.26	34.28	0.66
eng	mni	24_8_512_4096_2	14.54	32.52	0.65
eng	mni	24_8_1024_256_2	17.72	40.74	0.67
eng	mni	24_24_256_1024_2	3.18	12.24	0.4
eng	mni	24_24_512_2048_2	2.8	12.81	0.4
eng	mni	24_32_256_512_2	2.7	11.38	0.38
eng	mni	32_2_256_2048_2	7.27	24.19	0.61
eng	mni	32_2_512_4096_2	2.29	13.55	0.51
eng	mni	32_2_1024_256_2	17.45	39.5	0.67
eng	mni	32_12_256_1024_2	12.51	29.79	0.6
eng	mni	32_16_256_1024_2	15.71	33.21	0.6
eng	mni	32_16_512_2048_2	13.77	31.34	0.6
eng	mni	32_32_256_512_2	2.4	10.86	0.38
eng	mni	32_32_256_4096_2	1.27	8.05	0.37
eng	mni	32_32_512_1024_2	1.81	7.73	0.38