Causes and Cures for Interference in Multilingual Translation

Uri Shaham[τ] **Maha Elbayad**[μ] **Vedanuj Goswami**[μ]
Omer Levy[τμ] **Shruti Bhosale**[μ]

[τ] The Blavatnik School of Computer Science, Tel Aviv University
[μ] Meta AI

Abstract

Multilingual machine translation models can benefit from synergy between different language pairs, but also suffer from interference. While there is a growing number of sophisticated methods that aim to eliminate interference, our understanding of interference as a phenomenon is still limited. This work identifies the main factors that contribute to interference in multilingual machine translation. Through systematic experimentation, we find that interference (or synergy) are primarily determined by model size, data size, and the proportion of each language pair within the total dataset. We observe that substantial interference occurs mainly when the model is very small with respect to the available training data, and that using standard transformer configurations with less than one billion parameters largely alleviates interference and promotes synergy. Moreover, we show that tuning the sampling temperature to control the proportion of each language pair in the data is key to balancing the amount of interference between low and high resource language pairs effectively, and can lead to superior performance overall.

1 Introduction

Multilingual machine translation models can benefit from transfer between different language pairs (*synergy*), but may also suffer from *interference* (Ha et al., 2016; Firat et al., 2016; Aharoni et al., 2019; Arivazhagan et al., 2019). While there are methods to reduce interference and achieve better performance (Wang et al., 2020a; Kreutzer et al., 2021; Wang et al., 2021), such approaches are often compute intensive, and do not always work (Xin et al., 2022). In this work, we demonstrate that interference in multilingual translation largely occurs when the model is very small compared to the abundance of training data, and that the simple principled approach of enlarging the model and tuning the data sampling temperature provides a consistent solution to the interference problem that can even promote synergy.

This work methodically deduces the most simple ways of reducing interference in multilingual translation. We begin by inquiring what are the dominant factors that may interfere with learning to translate a particular language pair of focus $s \to t$, in the context of learning a multilingual translation model with many different language pairs. Controlled experiments show that besides model size and number of $s \to t$ training examples, the main factor that correlates with the level of interference is the proportion of *focus pair* examples ($s \to t$) observed out of the *total* number of examples (all language pairs) seen at each training step on average. Surprisingly, aspects like language similarity or number of translation directions have a much smaller effect.

In model and data scaling experiments, we observe that interference mainly occurs in extreme parameter poverty, when the language pair of focus is data-rich, but has to "share" a crowded parameter space with large quantities of other data. Enlarging the model to standard model sizes in machine translation literature alleviates interference and even facilitates synergy. For context, given a language pair of 15M sentence pairs that accounts for 20% of the total training data (75M), we observe severe levels of interference with 11M- and 44M-parameter transformers, but no interference when scaling the model to 176M parameters (the "big" model of Vaswani et al. (2017)) and significant synergy with 705M parameters. Interestingly, when the model is large enough, we find that increasing the amount of non-focus data to a certain point can further increase synergy.

Finally, given the evidence that data sizes and ratios strongly correlate with interference, we experiment with a natural lever that controls the proportion of each dataset in the overall mix in the simplest way: sampling temperature. Indeed, we

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15849–15863

find that calibrating the distribution of language pairs via temperature can substantially reduce the amount of interference in both high- and low-resource language pairs. Our results demonstrate the importance of tuning the temperature hyperparameter in multitask training, and suggest that previously reported accounts of severe interference in multilingual translation models might stem from suboptimal hyperparameter configurations.

2 Measuring Interference

We assume a common multilingual translation setup that involves L language pairs $s \rightarrow t$, where the source is always the same language s (English), and the target language t varies (English-to-many), or vice versa (many-to-English). The overall training data is a union of these training subsets, we note their sizes by $D_{s \rightarrow t}$. Sampling a training example x follows the distribution:

$$P(x \in s \rightarrow t) \propto \left(\frac{D_{s \rightarrow t}}{\sum_{s',t'} D_{s' \rightarrow t'}} \right)^{\frac{1}{T}} \quad (1)$$

Where T is the temperature hyperparameter (Devlin et al., 2019; Arivazhagan et al., 2019). $T = 1$ maintains the original data proportions, $0 < T < 1$ starves low resource language pairs, and $T > 1$ increases their representation in the training distribution. We mostly focus on the English-to-many setting in which interference is more apparent.[1]

We define interference as a negative interaction between different translation directions in a multilingual translation model. It is measured for a specific translation direction $s \rightarrow t$ by the relative difference in performance (test-set cross-entropy loss) between a bilingual model trained to translate only from s to t ($\mathcal{L}^{\text{bi}}_{s \rightarrow t}$) and a multilingual counterpart that is trained to translate other additional directions ($\mathcal{L}^{\text{multi}}_{s \rightarrow t}$):

$$\mathcal{I}_{s \rightarrow t} = \frac{\mathcal{L}^{\text{bi}}_{s \rightarrow t} - \mathcal{L}^{\text{multi}}_{s \rightarrow t}}{\mathcal{L}^{\text{bi}}_{s \rightarrow t}} \quad (2)$$

Negative values of $\mathcal{I}_{s \rightarrow t}$ indicate interference, while positive values indicate synergy.

3 Experimental Setup

Models We train encoder-decoder Transformer (Vaswani et al., 2017) models of 4 different sizes throughout our experiments. We use the original[2] transformer-base and transformer-big variants, as well as a smaller and a larger versions by adjusting the width of the architecture (Table 1).

Size	Hidden	FFN	Attn Heads	Params
XS	256	1024	4	11M
S	512	2048	8	44M
M	1024	4096	16	176M
L	2048	8192	32	704M

Table 1: Model sizes used in our experiments. Each model has 6 encoder and 6 decoder layers. We exclude the embeddings from the parameters count.

Data We use the multilingual benchmark introduced by Siddhant et al. (2020) based on WMT data. This benchmark includes a diverse set of 15 languages, each paired with English. The number of training examples is also diverse, ranging from 155K sentence pairs in Gujarati to 51M examples in Czech.[3] Table 2 provides additional dataset statistics.

Tokenization We build a shared vocabulary of 64K BPE tokens with sentencepiece (Kudo and Richardson, 2018) using a sampling temperature of 5 to increase the lower resource languages' representation. We use this vocabulary for all our experiments. We also add language ID tokens to our vocabulary, which are prepended to each source and target sequence to indicate the target language (Johnson et al., 2017).

Training We use Fairseq (Ott et al., 2019) to train transformer models with the Adam optimizer (Kingma and Ba, 2015) for up to 100K steps, with a dropout rate of 0.1, inverse square root learning rate schedule up to a maximum of 0.004, 8K warmup steps, and a batch size of 256K tokens. We choose the best checkpoint according to the average validation loss of all language pairs.

4 What Impacts Interference in Multilingual Translation?

We consider 5 factors that may potentially impact the performance of a given language pair $s \rightarrow t$ in the multilingual translation setting:

[1] Section 4.3 also includes many-to-English experiments, where we observe higher levels of synergy.

[2] With pre-layer normalization and a shared embedding matrix across the encoder input, decoder input, and decoder output (Press and Wolf, 2017).

[3] Note that Siddhant et al. (2020) only uses 11K pairs in Gujarati whereas we use the additional training data recommended by the WMT'19 shared task (https://statmt.org/wmt19/translation-task.html).

Language	ID	#Sentences (M)	Test Set
Czech	cs	51.769	WMT18
French	fr	40.853	WMT14
Russian	ru	38.492	WMT19
Chinese	zh	25.987	WMT19
Spanish	es	15.177	WMT13
Finnish	fi	6.587	WMT19
German	de	4.509	WMT14
Estonian	et	2.176	WMT18
Latvian	lv	0.638	WMT17
Lithuanian	lt	0.631	WMT19
Romanian	ro	0.610	WMT16
Hindi	hi	0.306	WMT14
Kazakh	kk	0.224	WMT19
Turkish	tr	0.207	WMT18
Gujarati	gu	0.156	WMT19

Table 2: Languages from the WMT-based benchmark of Siddhant et al. (2020), along with the number of sentence pairs in the training set, and the source of the test set. All languages are paired with English (en).

(1) Model size
(2) Training data size of $s \rightarrow t$, $D_{s \rightarrow t}$
(3) Proportion of $s \rightarrow t$ examples observed during training $P(x \in s \rightarrow t)$
(4) Total number of languages L
(5) Similarity between $s \rightarrow t$ and other pairs[4]

In the experiments we describe next, we provide empirical evidence that indicate the last two factors do not actually have a significant effect on the level of interference, and can therefore be pruned away. Subsequent experiments reveal that interference is indeed a function of model size, data size, and data proportion. Most striking is the fact that, across various data settings, enlarging the model to standard sizes consistently alleviates interference and may even promote synergy.

4.1 Does Language Similarity Matter?

Intuitively, data from languages that humans perceive as similar (e.g. languages that have some degree of mutual intelligibility, exhibit similar linguistic properties, or have shared vocabularies) should have a more positive effect on translation quality comparing to data from distinct languages (Lin et al., 2019; Wang et al., 2020b). To test this, we fix a *focus* language, and train *trilingual* models to translate from English to two languages, the focus language and an additional *interfering* language. We then look at interference trends as we vary the interfering language while controlling the amount of training data for each language pair.

Focus Language	#Examples	Other Language	#Examples
es	15.177M	fr⋆/cs/ru/zh	15.177M
es	0.118M	fr⋆/cs/ru/zh	15.177M
et	2.176M	fi⋆/fr/ru/zh	6.587M
et	0.118M	fi⋆/fr/ru/zh	6.587M

Table 3: Trilingual models for experiments on the impact of language similarity on interference. The most similar language to the focus language is noted with ⋆.

Setup We run two sets of experiments, one with Spanish (es, 15.2M parallel sentences) as the focus language, and another with Estonian (et, 2.2M examples). For each focus language, we select one of four interfering languages; Spanish is paired with French,[5] Czech, Russian, and Chinese, while Estonian is paired with Finnish,[6] French, Russian, and Chinese. To control the effects of data size in the English-Spanish experiments, we randomly sample 15.2M examples from each interfering language pair, making the ratio between focus and interfering languages 1:1. Similarly, in the English-Estonian experiments, we sample 6.6M examples from each interfering language to create a data ratio of 1:3. We also conduct similar experiments when we use only 118K focus language examples, to see the trends when the focus language pair is extremely low resource.[7] Table 3 provides an overview of the language similarity experiments.

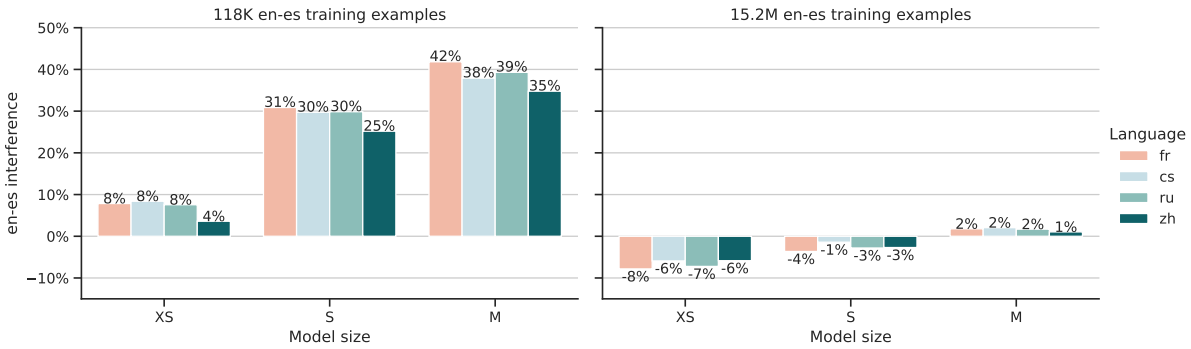
Results Figure 1a shows the interference rate for every model size when Spanish has only 118K parallel examples (left) and when using the full English-Spanish dataset (right). The variance in results somewhat correlates with language similarity when the dataset is very small, which aligns with previous work (Lin et al., 2019); French seems to help Spanish more than other languages when the model is big enough, while Chinese helps less. However, when training with the full dataset, the differences between other languages diminish for all model sizes. Concurrently, Fernandes et al. (2023) also found no significant difference for using French or Chinese as a third language combined with English-German in a very high resource

[4]While the other factors can be exactly quantified, it is not immediately clear how to measure language similarity. In our experiments, we use a phylogenetic interpretation of language similarity within the set of languages available in our dataset.

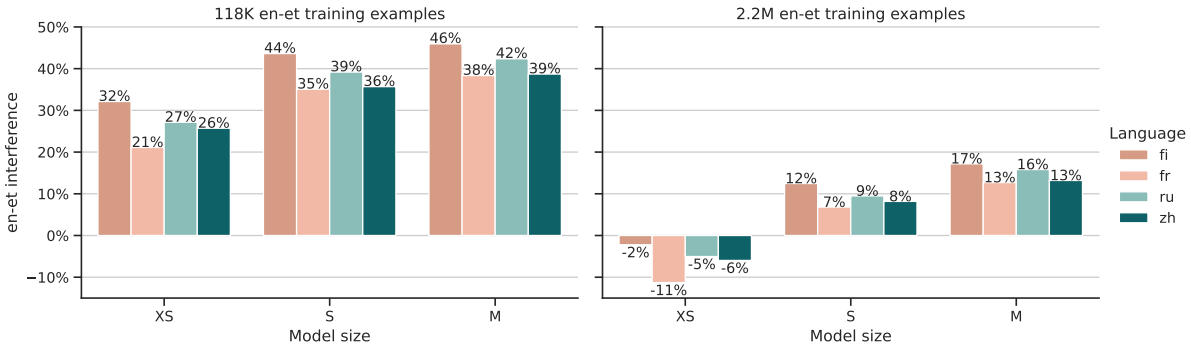
[5]Spanish and French are Western Romance languages.

[6]Estonian and Finnish are Balto-Finnic languages.

[7]118K sentence pairs is 128th of the English-Spanish training set. It is approximately equivalent to translating 30 novels.



(a) Models trained with 118K (left) and 15.2M (right) en-es examples together with 15.2M examples of en-xx.



(b) Models trained with 118K (left) and 2.2M (right) en-et examples together with 6.6M examples of en-xx.

Figure 1: Interference of models trained with en-es (a) or en-et (b) as low resource languages (left) and using their full training sets (right) together with one other language. Positive values indicate synergy, i.e. the focus language (es/et) loss of a trilingual model is lower (better) compared to its bilingual model baseline. Similarly, negative values indicate interference.

setting (600M examples per language pair).

We observe similar trends when Estonian is the focus language. Figure 1b shows that when Estonian only has 118K training examples, combining with Finnish data seems to have some positive effect. However, this effect also shrinks when using all of the English-Estonian train set (only 2.2M examples, compared to the 15.2M of English-Spanish) and a model that is not too small.[8]

4.2 Does the Number of Languages Matter?

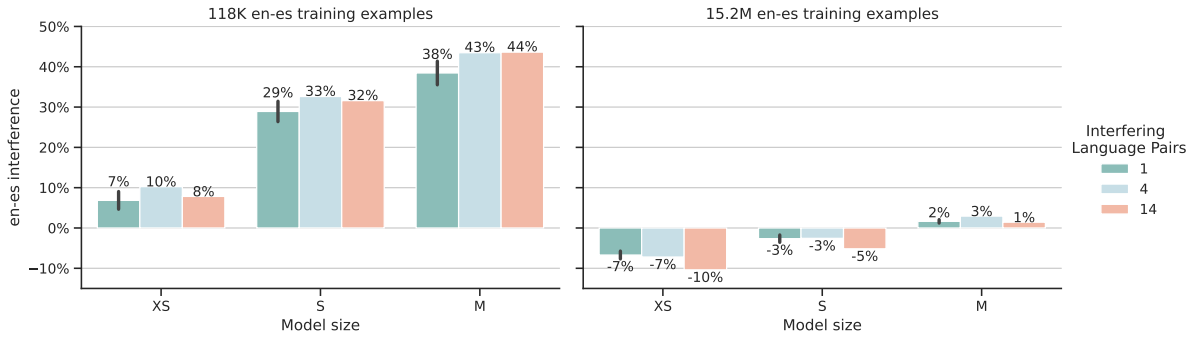
Do we get more interference when training with one interfering language pair or fourteen? We train models with varying numbers of language pairs while controlling for the overall number of interfering examples. We find that splitting the interfering data across more language pairs has a mild positive effect, which diminishes as the amount of focus-language data and/or model parameters scales up.

[8]See Figure 5 in Appendix A for the results of these experiments with absolute BLEU scores.

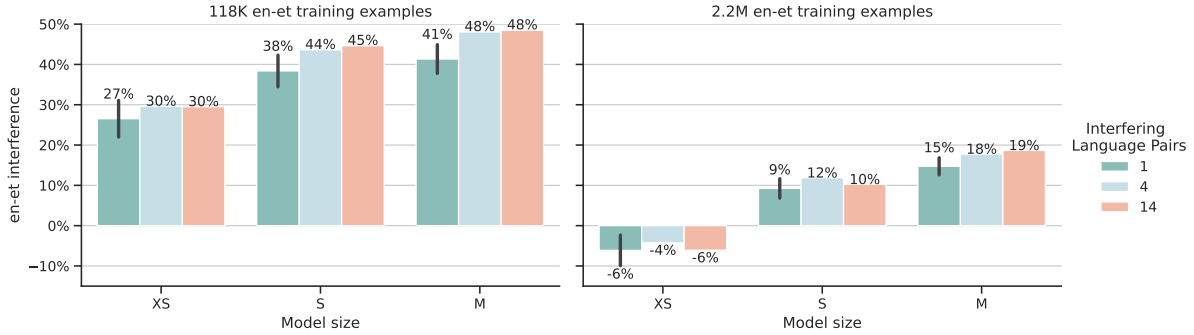
Focus Language	#Examples	Other Languages	#Examples
es	15.177M	cs/fr/ru/zh	15.177M
		cs+fr+ru+zh	15.177M
		cs+...+gu (14)	15.177M
es	0.118M	cs/fr/ru/zh	15.177M
		cs+fr+ru+zh	15.177M
		cs+...+gu (14)	15.177M
et	2.176M	fi/fr/ru/zh	6.587M
		fi+fr+ru+zh	6.587M
		cs+...+gu (14)	6.587M
et	0.118M	fi/fr/ru/zh	6.587M
		fi+fr+ru+zh	6.587M
		cs+...+gu (14)	6.587M

Table 4: Multilingual models for experiments on the impact of the number of other languages on interference. The trilingual model results are the average per focus language from Table 3.

Setup We train multilingual models on English-Spanish data alongside English to 1, 4, or 14 interfering languages. The interfering data always sums



(a) Models trained with 118K (left) and 15.2M (right) en-es training examples and 15.2M training examples for non-es languages.



(b) Models trained with 118K (left) and 2.2M (right) en-et training examples and 6.6M training examples for non-et languages.

Figure 2: en-es (a) and en-et (b) test interference of models trained with es (a) or et (b) as low resource languages (left) and using their full train sets (right) together with increasing number of languages, sharing a fixed budget of training examples. Positive values indicate synergy, i.e the focus language (es/et) loss of a multilingual model is lower (better) comparing to its bilingual model baseline. Similarly, negative values indicate interference.

up to a fixed 15.2M examples budget, distributed as evenly as possible among the different languages.[9] We repeat these experiments when Estonian is the focus language and the interfering example budget is 6.6M. Table 4 provides an overview of these experiments.

Results Figure 2a shows that more than one interfering language pair somewhat helps when English-Spanish has few training examples, but this effect largely disappears in the full training set and with larger models. We see similar trends for Estonian in Figure 2b, even though its full training set has only 2.2M examples. This phenomenon might be related to the fact that when the data distribution is sharp (i.e. one high resource paired with one very low resource) there is not enough incentive for the model to pay attention to the focus language's identifier token, compared to when the distribution is much more uniform. This result also corroborates similar findings for pretrained multilingual models (Conneau et al., 2020), although those experiments did not control the total quantity of data as in ours.[10]

4.3 The Impact of Model and Data Size

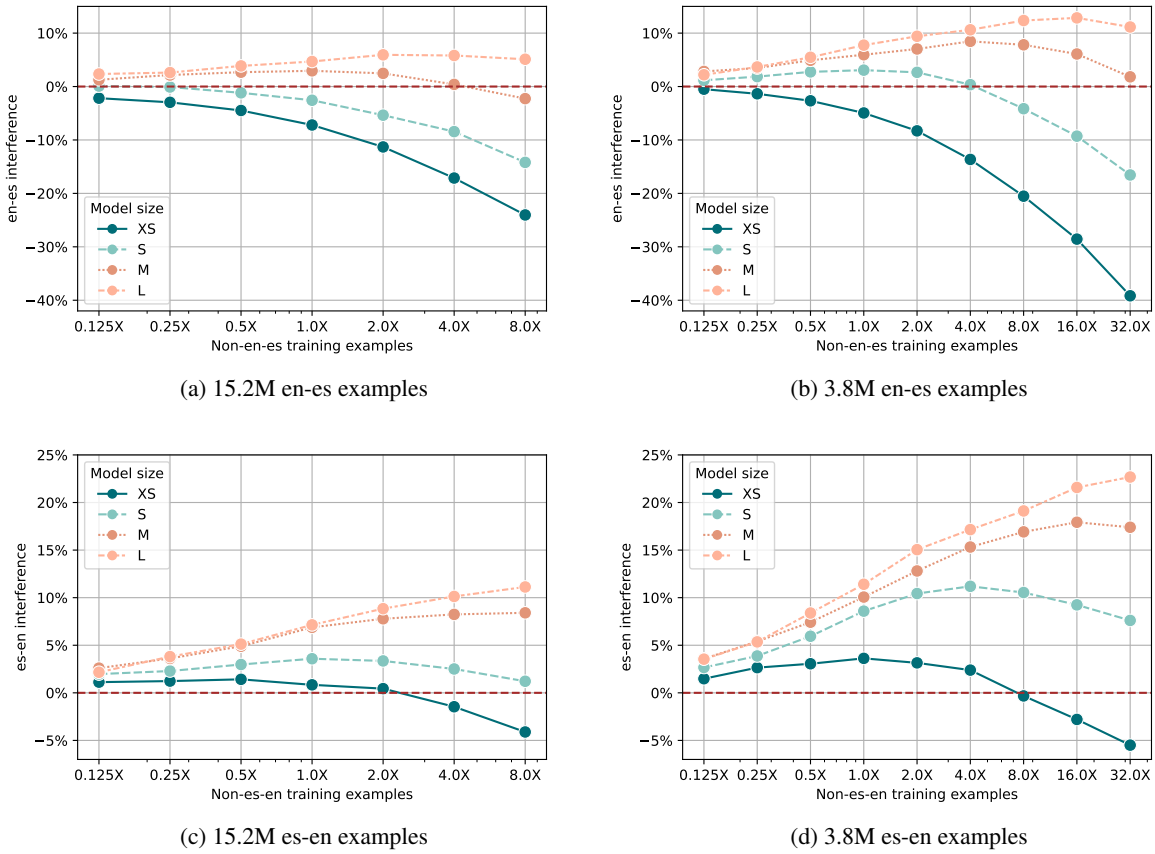
Seeing that language similarity and the number of interfering languages have only a limited effect on interference, we design a controlled setup to measure interference as a function of the remaining three factors: model size, focus language data size, and its proportion in the total amount of data seen during training.

Setup We train models using all the available 15.2M English-Spanish examples, with an increasing example budget for interfering language pairs, ranging from $1/8$ (1.9M) to 8 times (122M) the English-Spanish data, divided as evenly as possible between French, Czech, Russian, and Chinese.[11] To observe trends across $D_{s \to t}$ sizes, we

[9] Some languages have less than $15.2\text{M}/14$ (1.08M) examples. We use all of their training data, and divide the remaining budget evenly.

[10] See Figure 6 in Appendix A for the results of these experiments with absolute BLEU scores.

[11] Since Chinese has only 26M examples (less than $122\text{M}/4$), we use all of its train set in the 122M (8.0X) case,



(a) 15.2M en-es examples

(b) 3.8M en-es examples

(c) 15.2M es-en examples

(d) 3.8M es-en examples

Figure 3: Interference of en-es (top) and es-en (bottom) models trained using the full 15.2M en-es train set (left), and a sample of 3.8M en-es (right). Positive values indicate synergy, i.e. en-es or es-en loss of a multilingual model is lower (better) comparing to its bilingual model baseline. Similarly, negative values indicate interference.

rerun these experiments with a quarter (3.8M) of the English-Spanish data, while keeping the ratios with the rest of the data similar. Finally, we also conduct these experiments in the many-to-English setting.

Results Figures 3a and 3b show the interference and synergy for English-Spanish using a varying number of interfering examples. For smaller models (XS and S), increasing the amount of interfering data (i.e. decreasing the proportion of focus data) exacerbates interference. However, larger models appear to benefit from significant quantities of interfering examples; for instance, when training with $D_{s \to t} = 3.8$M, a large model (L) can gain over 10% relative loss improvement when there is 32 times more interfering data than focus data ($P(x \in s \to t) \approx 3\%$). Interestingly, we also observe that interference is sensitive to the ratio between model parameters and focus data, as the M model trained on 15.2M focus examples produces a similar curve to that of the 4-times smaller S model trained on 3.8M examples, both intersecting the synergy/interference line at the same point. Finally, Figures 3c and 3d show that when translating *into* English, interference is much less of an issue, occurring only in the XS model when the total amount of training data significantly exceeds the model's capacity. Scaling up the model not only improves the absolute performance (Appendix A), but also introduces substantial gains from synergy. Our results align with trends observed on cross lingual transfer when scaling pretrained multilingual models to 3.5 and 10 billion parameters (Goyal et al., 2021).

4.4 Tuning Interference with Temperature

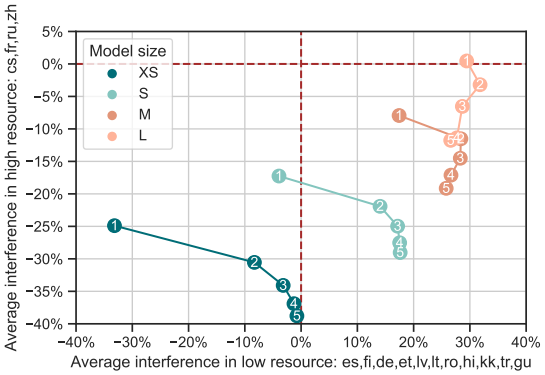
In the previous sections we demonstrated that the dominant factors impacting interference are the model size, the amount of focus language pair data $D_{s \to t}$, and the proportion of focus pair examples observed during training $P(x \in s \to t)$. In a

and sample the remainder of the example budget from the three from French, Czech, and Russian.

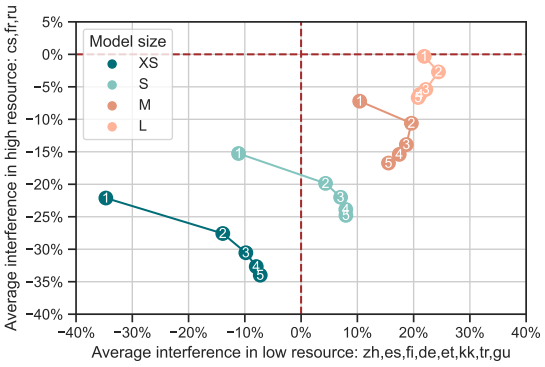
practical situation where both model size and multilingual data are fixed, how can one control the level of interference? Recalling Equation 1, we observe that the proportion of focus pair examples $P(x \in s \rightarrow t)$ is controlled via the temperature hyperparameter T. Although previous literature has largely used a value of $T = 5$ following Arivazhagan et al. (2019), our systematic experiments with different temperatures across three different data distributions and four model sizes suggest that this value can be sub-optimal and induce a substantial amount of interference, especially for model sizes that alleviate significant amounts of interference (M and L). Conversely, tuning the temperature shows that lower values ($T = 1, 2$) are typically able to reduce high-resource interference without harming low-resource synergy in our standard multilingual translation setting.

Setup We train models of four sizes with temperature ranging from 1 to 5 on three training distributions: (1) all available training data, (2) discarding 3 high resource languages (Czech, French and Russian), (3) discarding 4 low resource languages (Latvian, Lithuanian, Romanian and Hindi). When illustrating the results, we assign languages to high and low resource according to whether their relative data proportion decreases or increases when going from $T = 1$ to $T = 2$.

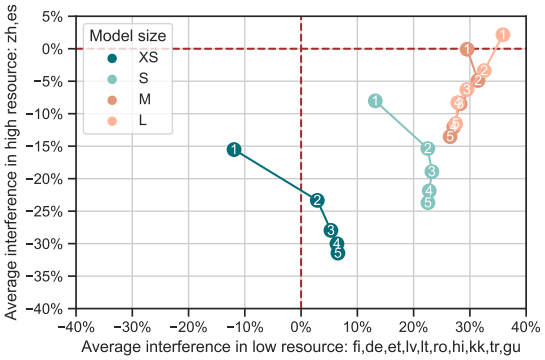
Results Figure 4 shows the trade-offs between the lower and higher resource languages, as defined above. First, we can see a clear trade-off for the smaller models (XS and S) from $T = 1$ to $T = 4$ in most cases. Increasing T helps promote synergy for low resource languages at the cost of increasing interference for the high resource languages. However, the larger models (M and L) clearly degrade when using $T \geq 3$; in fact, values of $T = 1$ and $T = 2$ are often better for high- and low-resource language pairs than the commonly-used $T = 5$. These results align with recent work Xin et al. (2022) showing that tuned scalarization is key to achieving strong bilingual baselines that often outperform more complicated multitask optimization methods.[12]



(a) Trained on all languages



(b) Trained w/o 4 low resource languages



(c) Trained w/o 3 high resource languages

Figure 4: Average interference/synergy of high (proportion declining when incrementing T) and low (proportion ascending when incrementing T) resource languages of different model sizes (colors) for different training distributions (a,b,c) using T values ranging from 1 to 5 (numbers on markers). Positive values indicate synergy and negative values indicate interference.

5 Related Work

Scaling Laws in Machine Translation Previous work also looked at scaling trends of data and models sizes for machine translation. Gordon et al. (2021) proposed scaling laws in the data and model parameters and demonstrated their ability to predict the validation loss of bilingual translation models from Russian, Chinese, and German to English. Ghorbani et al. (2022) found scaling laws for different configurations for the encoder and decoder,

[12] See Table 5 in Appendix A for the results of these experiments with absolute BLEU scores.

independently varying the number of layers in each of them. Bansal et al. (2022) examined different architectures and described data size scaling laws for machine translation in a large scale for English to German and English to Chinese. While all of these works focused on the bilingual setting, we unveil trends for multilingual translation, which has increased complexity. Concurrently to our work, Fernandes et al. (2023) proposed scaling laws for multilingual machine translation, focusing on trilingual models trained on English-German with English-Chinese or French

Multitask Methods for Multilingual Machine Translation Multitask methods have been proposed extensively to enhance the performance of multilingual translation models. Some utilize validation based signals to determine which language pairs should be prioritized throughout training, either with adaptive scheduling (Jean et al., 2019), gradient similarities to the validation set Wang et al. (2020a), or a multi-armed bandits model (Kreutzer et al., 2021). Zhu et al. (2021) added dedicated embedding and layer adapter modules to the Transformer, and Lin et al. (2021) suggested learning a binary mask for every model parameter and every language pair, both requiring further training after the base multilingual model converges. Li and Gong (2021) used per language gradients geometry to rescale gradients of different language pair to improve performance on low resource languages. Wang et al. (2021) extended PCGrad (Yu et al., 2020) to create Gradient Vaccine, a method that attempts to deconflict different language pairs gradients by replacing them with more similar vectors in terms of cosine similarity. While the motivation for these methods is clear and intuitive, they are usually more complex and computationally expensive than the baseline. Moreover, their efficacy is often demonstrated using relatively small[13] models, while modestly increasing the model size can both strengthen the bilingual baselines and reduce the interference problem significantly.

Critical Takes on Multitask Optimization Methods Multitask optimization methods were recently under scrutiny. Kurin et al. (2022) experimented with many of those for image classification and reinforcement learning problems, and found that none of them consistently outperformed a well tuned baseline with proper use of known regularization techniques. Similarly, Xin et al. (2022) showed that despite their increased complexity, no popular multitask method was superior to a sweep over scalarization weights for a baseline trilingual translation model. This work complements this line of research by examining *multilingual* translation models and how can modest scale and calibrated temperature reduce problems associated with multitasking.

[13]Transformer-base or big from Vaswani et al. (2017).

6 Conclusion

This work examines the dominant factors that influence interference in multilingual machine translation. Namely, the model size, the amount of parallel data for the focus language pair, and the proportion of examples from the focus language pair with respect to the total data seen during training. While specialized multitask techniques are sometimes demonstrated on small transformer models, we find that a standard baseline model of 176M parameters reduces the interference problem significantly, and further scaling up results in synergy among the different language pairs. We further demonstrate the importance of tuning the temperature at which different language pairs are sampled during training; while existing literature largely relies on high temperatures, which indeed improve low-resource performance in parameter-poor settings, larger models benefit from a more natural distribution that reflects the raw training data. These simple strategies for addressing interference call into question the necessity and perhaps even the validity of recently-proposed complex anti-interference methods and reaffirm the tried-and-true method of increasing model capacity to accommodate for higher data diversity.

7 Limitations

One limitation of this work is the focus on English-to-many and many-to-English settings, while previous studies also went beyond English-centric translation (Freitag and Firat, 2020; Fan et al., 2022). Second, we experiment with a WMT based benchmark that has a total of 15 languages and 200M training examples, when translation models were also trained on larger datasets (Aharoni et al., 2019; Arivazhagan et al., 2019; NLLB Team et al., 2022). We leave questions about the amount of scale that will be required to effectively mitigate interference in massively (many-to-many, billions of parallel sequences) multilingual settings for future work.

Additionally, the data collected from high resource languages may be of higher quality compared to that collected from low resource languages. Further research is needed to determine the impact of low quality training data on interference and synergy. Finally, while we explore trends when scaling models width, deeper models (Ghorbani et al., 2022) might help mitigating interference even further.

Acknowledgments

This research is supported by the Yandex Initiative in Machine Learning. We thank Maor Ivgi, Yilin Yang, Jean Maillard, and Ves Stoyanov for their valuable feedback.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in NMT: The effect of noise and architecture. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1466–1482. PMLR.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2022. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 550–560, Online. 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, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning. *ArXiv*, abs/1909.06434.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's

multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Julia Kreutzer, David Vilar, and Artem Sokolov. 2021. Bandits don't follow rules: Balancing multi-facet machine translation with multi-armed bandits. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3190–3204, Punta Cana, Dominican Republic. Association for 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.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. 2022. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems*.

Xian Li and Hongyu Gong. 2021. Robust optimization for multilingual translation with imbalanced data. In *Advances in Neural Information Processing Systems*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online. Association for Computational Linguistics.

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

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.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020a. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020b. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 4438–4450, Online. Association for Computational Linguistics.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. 2021. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. 2022. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems*.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.

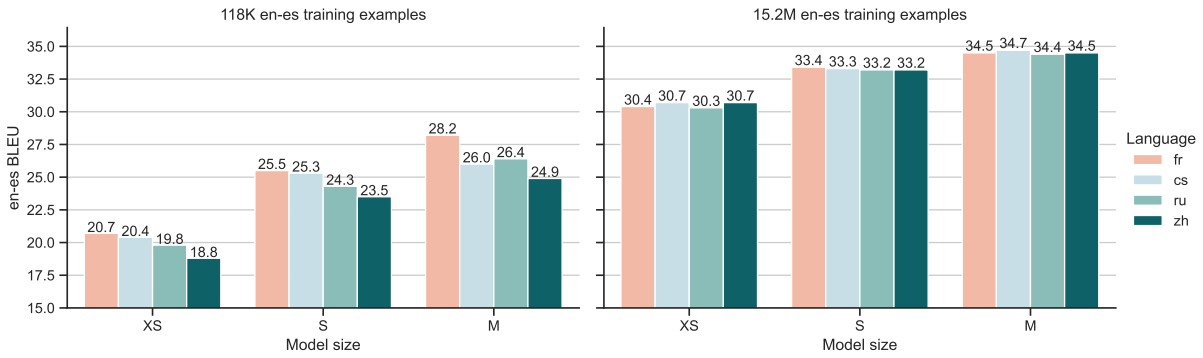
Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2812–2823, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A BLEU Scores

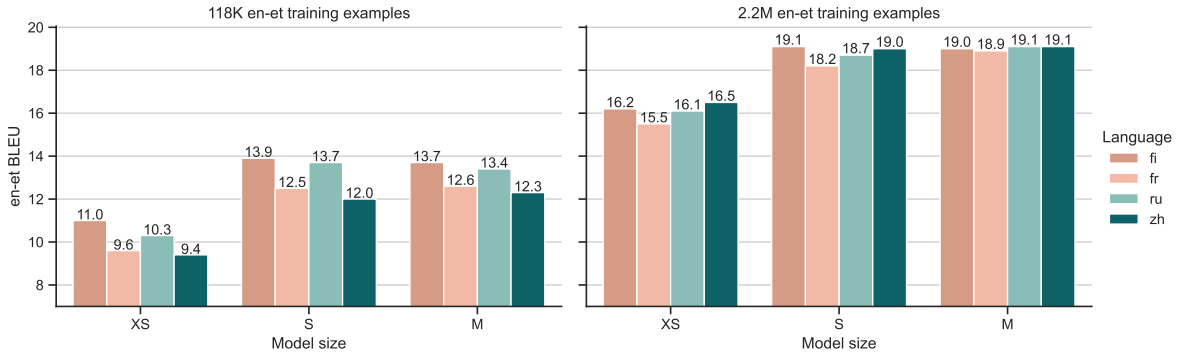
Throughout the paper we calculate interference in terms of test loss values. We additionally provide the test BLEU scores achieved by our models. We generate using beam search with 5 beams, without length penalty. We use SacreBLEU (Post, 2018) to calculate test sets BLEU (Papineni et al., 2002) scores.

Language similarities Figure 5 shows BLEU scores of models from experiments in Section 4.1. They reflect similar trends, as the variance between different interfering languages when the focus language has only 118K examples diminish when a decent amount of training data is available.

Number of languages Figure 6 shows BLEU scores of models from experiments in Section 4.2. They also demonstrate that low resource pairs benefit when there are more interfering languages, but this effect disapper with a decent amount of training data.

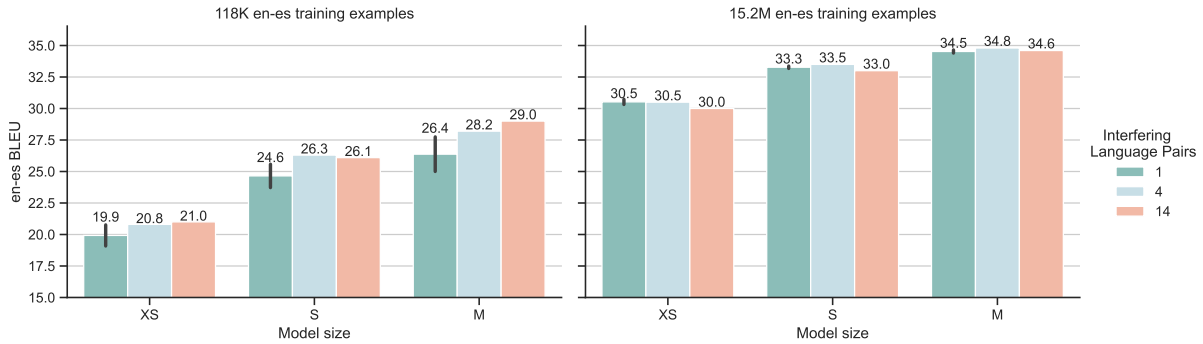


(a) Models trained with 118K (left) and 15.2M (right) en-es training examples and 15.2M training examples for non-es languages.

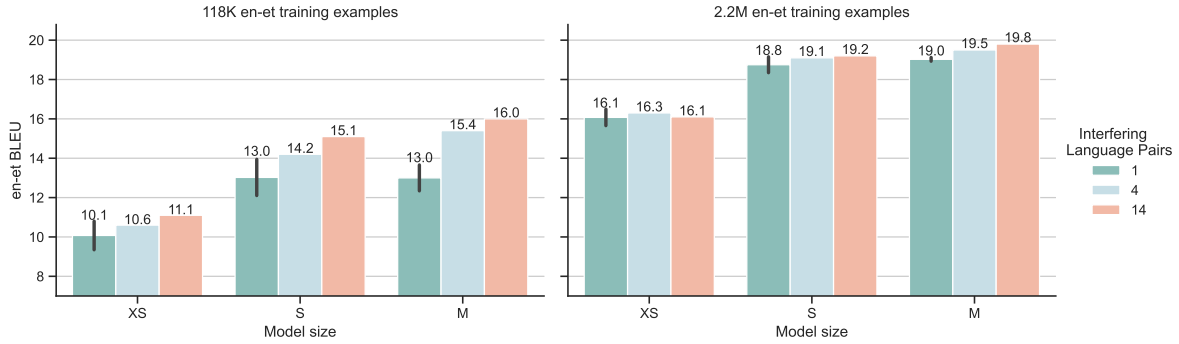


(b) Models trained with 118K (left) and 2.2M (right) en-et training examples and 6.6M training examples for non-et languages.

Figure 5: en-es (a) and en-et (b) test BLEU scores of models trained with es or et as low resource languages and using their full train sets together with one other en-xx pair.



(a) Models trained with 118K (left) and 15.2M (right) en-es training examples and 15.2M training examples for non-es languages.



(b) Models trained with 118K (left) and 2.2M (right) en-et training examples and 6.6M training examples for non-et languages.

Figure 6: en-es (a) and en-et (b) test BLEU scores of models trained with es or et as low resource languages and using their full train sets together with increasing number of languages, sharing a fixed budget of training examples.

Size	Tmp	cs	fr	ru	zh	es	fi	de	et	lv	lt	ro	hi	kk	tr	gu
XS	bi	**19.6**	**35.1**	**24.2**	**27.3**	**31.7**	**17.7**	**24.1**	**17.5**	12.1	9.2	**22.4**	6.5	0.5	7.7	1.6
	1	16.7	31.9	20.5	20.8	27.4	12.8	15.7	10.4	8.0	6.1	15.6	4.8	1.0	4.9	2.5
	2	15.5	30.8	19.0	20.3	27.4	14.1	17.6	13.1	11.3	9.4	20.4	9.1	2.3	9.1	6.4
	3	15.2	30.2	18.6	19.6	27.3	14.6	18.0	13.5	11.9	9.8	21.5	11.0	3.2	10.2	7.6
	4	14.8	30.1	18.2	19.4	27.1	14.7	18.1	13.7	12.4	10.0	21.5	11.7	3.2	10.7	8.7
	5	14.5	29.9	17.6	19.0	27.1	14.6	18.1	13.6	**12.6**	**10.3**	21.7	**11.9**	**3.5**	**10.8**	**9.2**
S	bi	**22.1**	**38.4**	**27.2**	**29.9**	**33.8**	**19.8**	**26.1**	17.4	12.0	8.5	22.1	4.8	0.5	7.2	1.8
	1	20.3	36.2	24.7	26.4	31.0	16.8	20.8	14.5	12.1	9.8	21.0	7.9	1.7	7.6	4.9
	2	19.9	35.7	24.1	25.7	31.4	18.5	22.4	17.3	14.9	12.2	24.1	14.1	4.6	12.1	11.6
	3	19.2	35.6	23.5	25.6	31.2	18.4	22.5	17.6	15.3	12.5	24.5	15.2	5.6	12.9	13.1
	4	19.1	35.2	23.7	25.0	30.9	17.5	22.6	**17.7**	**15.4**	**12.8**	**25.0**	15.3	5.7	13.3	13.1
	5	18.5	34.8	23.4	25.1	30.9	18.1	22.3	17.5	15.3	12.5	24.9	**15.4**	**5.9**	**13.8**	**13.5**
M	bi	**23.1**	**40.1**	**28.8**	**30.7**	**34.2**	19.6	25.9	17.1	11.5	7.8	21.6	4.0	0.4	5.9	1.0
	1	22.4	39.6	27.3	29.8	33.6	19.1	24.1	18.0	14.6	12.0	23.9	12.4	3.6	10.7	8.2
	2	22.1	39.3	26.5	29.7	33.5	19.5	25.7	19.3	17.1	13.8	**26.5**	**15.9**	**6.3**	14.1	**14.2**
	3	21.8	38.0	26.1	29.6	33.4	20.1	**26.1**	**20.2**	**17.4**	13.8	**26.5**	15.2	5.8	**14.2**	14.1
	4	21.3	38.0	25.9	29.0	33.4	**20.3**	25.8	20.1	16.9	**14.1**	**26.5**	14.6	5.5	13.7	12.2
	5	21.1	37.7	26.2	28.6	32.8	19.9	25.6	19.4	16.8	13.9	26.3	14.6	5.2	13.8	12.3
L	bi	22.9	40.0	28.5	30.7	34.4	18.6	25.8	16.9	10.8	8.5	21.4	3.8	0.4	5.4	1.3
	1	**23.4**	**40.7**	**29.4**	**31.4**	34.8	20.7	26.5	19.2	16.3	13.4	26.1	**14.4**	4.6	12.5	10.3
	2	23.0	40.4	29.1	31.1	34.7	20.6	**28.0**	20.2	**17.9**	**14.2**	**26.7**	14.2	**4.7**	**14.2**	12.4
	3	22.9	39.8	28.4	31.1	**34.9**	**21.3**	27.7	**20.5**	17.4	**14.2**	26.2	13.5	4.6	14.0	12.2
	4	22.1	39.2	26.5	29.8	34.0	20.5	26.7	20.3	17.3	**14.2**	26.4	13.8	**4.7**	14.0	12.1
	5	21.9	38.9	27.5	30.1	34.1	21.1	26.7	20.4	17.2	13.7	25.8	13.6	3.8	13.9	**13.0**

Table 5: Test BLEU scores across four model sizes of bilingual baselines (bi) and multilingual models trained with temperature values $T \in [1, 5]$.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☒ A2. Did you discuss any potential risks of your work?
Our work does not add new risks involving translation models

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Sections 0,1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The WMT data used in our experiment is a common machine translation dataset and is publicly available research purposes.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The usage was consistant with the artifacts intended use.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The WMT data used in our experiment is a common machine translation dataset and is publicly available research purposes.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Mostly languages in section 3. Regarding the rest, adding justification from above: The WMT data used in our experiment is a common machine translation dataset and is publicly available research purposes.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.