Scaling Laws for Multilingual Neural Machine Translation

Patrick Fernandes [1 2 3] **Behrooz Ghorbani** [1] **Xavier Garcia** [1] **Markus Freitag** [1] **Orhan Firat** [1]

Abstract

In this work, we provide a large-scale empirical study of the scaling properties of multilingual neural machine translation models. We examine how increases in the model size affect the model performance and investigate the role of the training mixture composition on the scaling behavior. We find that changing the weightings of the individual language pairs in the training mixture only affect the multiplicative factor of the scaling law. In particular, we observe that multilingual models trained using different mixing rates all exhibit the same scaling exponent. Through a novel joint scaling law formulation, we compute the *effective number of parameters* allocated to each language pair and examine the role of language similarity in the scaling behavior of our models. We find little evidence that language similarity has any impact. In contrast, the direction of the multilinguality plays a significant role, with models translating from multiple languages *into* English having a larger number of effective parameters per task than their reversed counterparts. Finally, we leverage our observations to predict the performance of multilingual models trained with *any* language weighting at *any* scale, significantly reducing efforts required for language balancing in large multilingual models. Our findings apply to both in-domain and out-of-domain test sets and to multiple evaluation metrics, such as ChrF and BLEURT.

1. Introduction

Over the past few years, scaling has emerged as a popular and effective way to improve the performance of neural networks (Brown et al., 2020; Chowdhery et al., 2022; Lepikhin et al., 2020). Given the costs associated with training large neural models, much work has gone into understanding their scaling properties and predicting the evolution of their performance with scale through **scaling laws**. Such scaling laws have been instrumental in guiding the model development efforts across a variety of domains such as computer vision (Zhai et al., 2022), language modeling (Kaplan et al., 2020; Hoffmann et al., 2022), and neural machine translation (Ghorbani et al., 2022).

Despite these impressive developments, most of the scaling laws studies available in the literature only focus on single-task, single-language models. On the contrary, current massive neural models are often trained to solve more than one task across one or more modalities & languages (Chowdhery et al., 2022; Sanh et al., 2022; Reed et al., 2022). This disconnect from the current research frontier limits the applicability of scaling laws in guiding model development decisions. In particular, currently available scaling laws studies are unable to inform the decision process on **balancing the different tasks effectively** at training time. Without such guidance, practitioners often have to rely on cumbersome and costly approaches such as approximate grid search to inform their decision-making; such approaches quickly become infeasible as the problem scale grows.

In this paper, we take the initial step towards developing a quantitative understanding of the scaling behavior for multitask models. We choose multilingual neural machine translation (MNMT) as the setup for this initial study. This choice is motivated by several reasons: (1) MNMT has been framed and studied as a multi-task optimization problem extensively in the past (Dong et al., 2015; Luong et al., 2015; Arivazhagan et al., 2019b; Wang et al., 2021); (2) It provides a popular setup with mature benchmarks and substantial literature on scaling (Lepikhin et al., 2020; Costa-jussà et al., 2022; Bapna et al., 2022; Huang et al., 2019); (3) Moreover, recent results on scaling laws for single-task MT models provide a natural starting point for our study (Ghorbani et al., 2022; Bansal et al., 2022; Gordon et al., 2021; Zhang et al., 2022). (4) Finally, recent findings on the optimization dynamics of MNMT models greatly simplify our study by removing the need to examine the role of the optimization algorithm in our results (Xin et al., 2022).

For our analysis, we train over 200 MNMT models (ranging

[1]Google Research [2]Carnegie Mellon University [3]Instituto Superior Técnico. Correspondence to: Patrick Fernandes <pfernand@cs.cmu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

from 20M to 1B non-embedding parameters) and systematically examine their scaling behaviors. We focus our investigation on the **data-rich, compute-rich regime** where we have access to vast amounts of training data for all the language pairs (i.e. tasks)[1] and the models are trained to near convergence. Here, the main bottleneck in the model performance is the lack of model capacity. We establish the following observations:

- For each fixed training mixture, the evolution of the test cross-entropy loss for the i_{th} language pair ($\mathcal{L}_i$) with model size (N) follows a scaling law that resembles the scaling behavior of single-language-pair models:

$$\mathcal{L}_i(N; \boldsymbol{p}) \approx \beta_{\boldsymbol{p},i} N^{-\alpha_{\boldsymbol{p},i}} + L_{\infty}^{(\boldsymbol{p},i)}. \quad (1)$$

  Here, $\boldsymbol{p}$ is a vector of probabilities that determines the weight of each language pair in the training mixture. Furthermore, we find that changes in the language pair weightings only affect the multiplicative factor β; the scaling exponent α and the irreducible loss L_∞ are unaffected by these changes. As such, our results suggest that scaling multilingual models improves the loss at rates that are independent of the weights of the individual language pairs in the training mixture.

- We leverage these findings to propose a scaling law that jointly predicts the performance for all language pairs and weightings considered, and use it to examine how the model splits its capacity in between the language pairs by computing the **effective number of parameters** allocated to each language pair (Section 3.3).

- We examine the popular belief that training multilingual models on similar languages is more effective than training models in unrelated languages. Surprisingly, for the high-resource language pairs considered, we do not observe any significant differences in the scaling behavior of models trained to translate from English into related languages (En→{De, Fr}) and models trained in unrelated languages (En→{De, Zh}). In contrast, we observe that models trained to translate from multiple languages into English (XX→En) benefit much more from multitasking compared to those trained on translation out of English (En→XX).

- By approximating the capacity splitting behavior of multilingual models, in Section 3.4, we provide a scaling law that predicts **the full task performance trade-off frontier** as a function of the model size N (See Figure 7). In Section 3.4, we describe how such predictions can be leveraged for efficiently guiding task balancing when training large multilingual models.

[1]Using machine translation terminology, all language pairs are *high-resource*.

2. Background

2.1. Neural Scaling Laws

Recent research suggests that the performance of large neural models is well-predicted by a smooth function of the fundamental problem parameters: the model size N [2], the size of the training data D, and the amount of compute used for training C (Hestness et al., 2017; Rosenfeld et al., 2019; Kaplan et al., 2020; Hernandez et al., 2021). The most relevant of these studies to ours is Ghorbani et al. (2022) where the authors study the effects of increasing the model size for single-task NMT models in the data-rich ($D \to \infty$), compute-rich ($C \to \infty$) regime. In this setting, the authors show that the following *bivariate* law describes the scaling behavior of encoder-decoder Transformers

$$\mathcal{L}(N_e, N_d) = \beta N_e^{-\alpha_e} N_d^{-\alpha_d} + L_\infty. \quad (2)$$

Here, N_e and N_d correspond to the number of parameters in the encoder and decoder respectively and L_∞ corresponds to the irreducible loss associated with the task. $\{\beta, \alpha_e, \alpha_d, L_\infty\}$ are the parameters of the scaling law that need to be empirically estimated from the data.

In addition, (Ghorbani et al., 2022) examine the question of optimally allocating parameters between the encoder and the decoder. They show that in order to attain the optimal scaling behavior, one needs to proportionally scale the encoder and the decoder together. Under such scaling scheme, Equation 2 simplifies to

$$\mathcal{L}(N) = \beta N^{-\alpha} + L_\infty, \quad (3)$$

which is similar to the scaling behavior observed in other domains such as computer vision (Zhai et al., 2022) and autoregressive generative models (Henighan et al., 2020).

Based on these results, to achieve the optimal scaling behavior, we adopt the proportional encoder-decoder scaling scheme for our experiments. A detailed overview of the size and architecture of our models is presented in Appendix A.

2.2. Multitask Optimization

Multilingual NMT is commonly framed and studied as a multitask optimization problem (Dong et al., 2015; Luong et al., 2015; Arivazhagan et al., 2019b; Wang et al., 2021).

We focus our investigation on the supervised learning setup where the model parameters $\boldsymbol{\theta} \in \mathbb{R}^N$ are trained on K different tasks simultaneously. In multilingual MT, each task corresponds to translation for a different language pair. We denote the loss associated with task i with $\mathcal{L}_i(\boldsymbol{\theta})$.

Multitask models are often trained by minimizing a convex

[2]Following the literature conventions, we only consider the non-embedding layers when computing N.

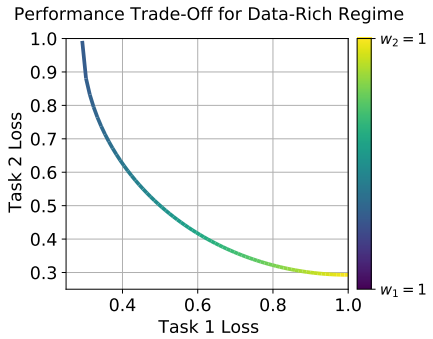

Figure 1. Cartoon representation of the performance trade-off frontier for a hypothetical model.

combination of the per-task losses:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{w}) = \arg\min \sum_{i=1}^{K} \boldsymbol{w}_i \mathcal{L}_i(\boldsymbol{\theta}) \; ; \; \boldsymbol{w} > 0, \; \sum_{i=0}^{K} \boldsymbol{w}_i = 1 \quad (4)$$

Here, $\boldsymbol{w}$ is a fixed vector of the task weights, determined apriori by the practitioner to emphasize her preferences on the balancing of the tasks. This so-called **scalarization** approach is highly popular in the community due to its effectiveness and simplicity.[3] In fact, despite this simplicity, recent results on multitask optimization suggest that scalarization achieves performances on par or better than bespoke optimizers designed specifically for multitask models (Xin et al., 2022; Kurin et al., 2022).

In current large text models, such explicit scalarization is rare. Instead, scalarization is often implemented **implicitly**, by sampling observations from each task proportionally to that task's weight. Proportional sampling produces (in expectation) the same overall loss function as explicit scalarization but with much less engineering complexity.

Xin et al. (2022) demonstrate that there exists a smooth, well-defined performance trade-off frontier for multitask models in the data rich regime. This frontier represents the performance trade-off the model is able to achieve in between the tasks as a function of the task weights (see Figure 1 for a cartoon representation). Naturally, finding an accurate characterization of the performance trade-off frontier is key in finding a systematic solution to the task balancing issue.

3. Effects of Scale in Multilingual MT

3.1. Experimental Setup

We use the pre-LN encoder-decoder Transformer architecture in our models (Xiong et al., 2020; Vaswani et al., 2017). We train models of up to 8 sizes, approximately ranging from 20M to 1B (non-embedding) parameters. When scaling encoder-decoder Transformers, to achieve the optimal scaling behavior, we scale the encoder and the decoder proportionally by increasing the model dimension and the number of layers in tandem. See Appendix A for details.

For our experiments, we train two cohorts of models: En→XX and XX→En. For En→XX cohort, we train multilingual model for translation from English to {German (De), Chinese (Zh)} and {German (De), French (Fr)}. For XX→En cohort, we present results for {De, Zh}→En.

We use the *implicit* scalarization approach to train our models; each observation in the training batch is chosen from the first language pair with probability p and the second language pair with probability $1 - p$. For our experiments, we choose p from the set

$$p \in \{0, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 1\}. \quad (5)$$

For En→XX models, to avoid confusing the model, we prepend a language token to the source sentence specifying the target language (e.g. `<2de>`). The models are trained with per-token cross-entropy loss and Adafactor optimizer (Shazeer & Stern, 2018), using a fixed batch size of 500K tokens and inverse square root learning rate schedule. To mirror the compute-rich regime as closely as possible, we trained our models to near convergence. In practice, this translates to training our smaller models ($<$ 500M parameters) for 500K gradient steps and larger models for 1M steps.

To place our models in the data-rich regime, we use a massive in-house web-crawled dataset for training our models. We filter this data using an online data selection procedure (Wang et al., 2018) and high-quality web-domain reference sets, extracting 600M sentences for each language pair in the En→XX direction and 1.2B sentences for the XX→En language pairs. We tokenize this corpus by using a pretrained multilingual SentencePiece (Kudo, 2018) vocabulary, with a size of 128K sub-words.

We measure the performance of models on both *in-domain* and *out-of-domain* test sets. For the in-domain test set, we extract 2000 sentences from the same in-house datasets used to create the training (ensuring no overlap). For out-of-domain, we use *newstest2019* (Barrault et al., 2019), consisting of 2000 sentence-pairs extracted from aligned news documents.

3.2. Results & Analysis

Understanding Multilingual Scaling We start our analysis by independently examining the model scaling behavior for each individual language pair weighting p in (5). For each choice of p, we fit a scaling law of the form

$$\mathcal{L}_i(N; p) = \beta_{p,i} N^{-\alpha_{p,i}} + L_{\infty}^{(p,i)} \quad (6)$$

[3] See (Boyd & Vandenberghe, 2004) for more a detailed discussion of scalarization.

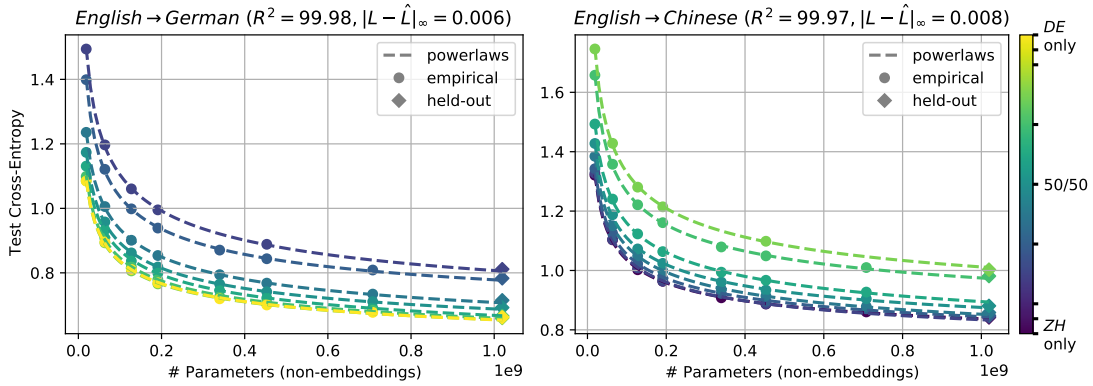

Figure 2. The evolution of the in-domain test cross-entropy loss with model size for En→{De, Zh} models, as well as the fitted scaling laws. These scaling laws are **fitted separately for each language pair weighting**. The color represents the weighting of the languages. The scaling laws are able to capture close to 100% of the variation in the data for both language pairs. Note that we don't show the *zero-shot* behavior.

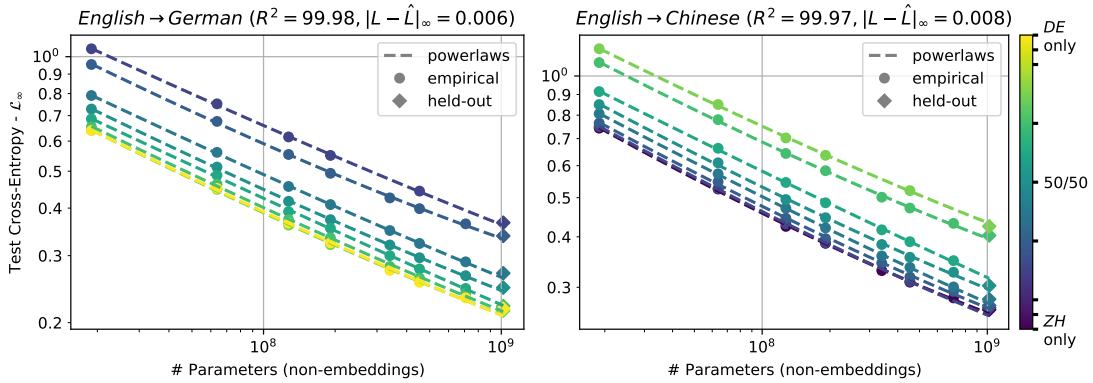

Figure 3. Log-log plot of the evolution of the (in-domain) test cross-entropy loss as we scale. We subtract a constant $L_\infty^{(i)}$, jointly fitted for all the weightings (Equation 7). All lines are nearly parallel, suggesting that the scaling exponent is unchanged for all p.

to the empirical (test) performance of models resulting from that language pair weighting.

Figure 2 presents our findings for En→{De, Zh} models. Each point on the graph corresponds to the empirical test-cross entropy performance of a model at the end of the training.[4] We observe that our per-weighting laws are able to capture the scaling behavior of our multilingual models on both language pairs. As expected, when the weight for one of the languages is decreased, the performance of the models on that language decreases for all scales. Our results suggest that the benefits of the increased model size for MNMT models are well-described by a power-law. See Appendix B for similar results for other language pair combinations.

Figure 4 shows the fitted scaling law coefficients for different values of p. The shaded area marks the one standard deviation uncertainty interval of our estimates.[5] Interestingly, we find that, across all values of p, both the scaling exponent (α) and the irreducible loss ($\mathcal{L}_\infty$) seem to be relatively unchanged. In particular, all of our estimated α and $\mathcal{L}_\infty$ parameters are within two standard deviations of each other. In contrast, the multiplicative factor β seems to be highly sensitive to the choice of p.

Figure 3 visually confirms the assertion that for our models α_p and L_∞ are effectively constant. Here, we have subtracted a fixed constant $L_\infty^{(i)}$ from all the Figure 4 curves corresponding to the language pair i. We then plot results on log-log axes. As the figure suggests, the lines are all near parallel, suggesting that the scaling exponent is unchanged for all p. In practical terms this means that, for example,

[4]For low probability language pairs, we apply a convergence correction procedure to make up for slow convergence. See Appendix G for more details.

[5]We gauge the uncertainty in the coefficients by measuring the fluctuations in our estimates when our empirical datapoints are perturbed by $\epsilon \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. We choose a conservative σ of 1% of the observed empirical loss for each data point.

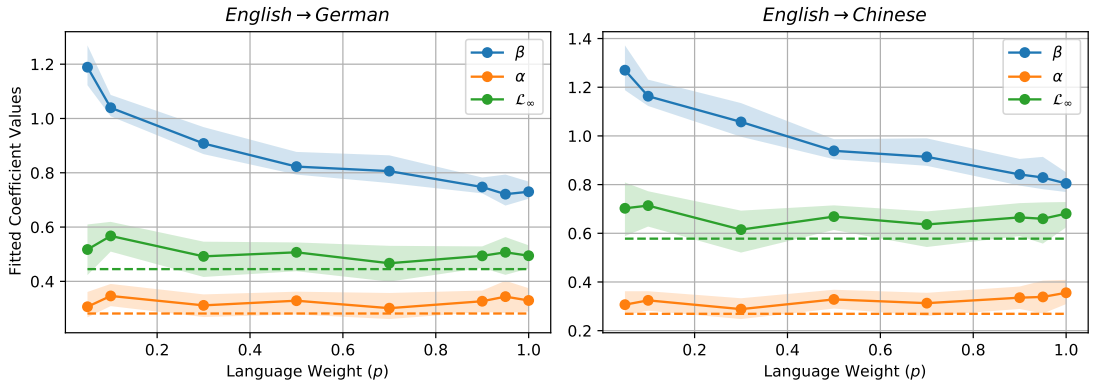

Figure 4. Coefficient values for German (left) and Chinese (right) as a function of the language weight, with the shaded region representing the standard deviation. The dashed lines represent the value of jointly fitted coefficients from Equation 7.

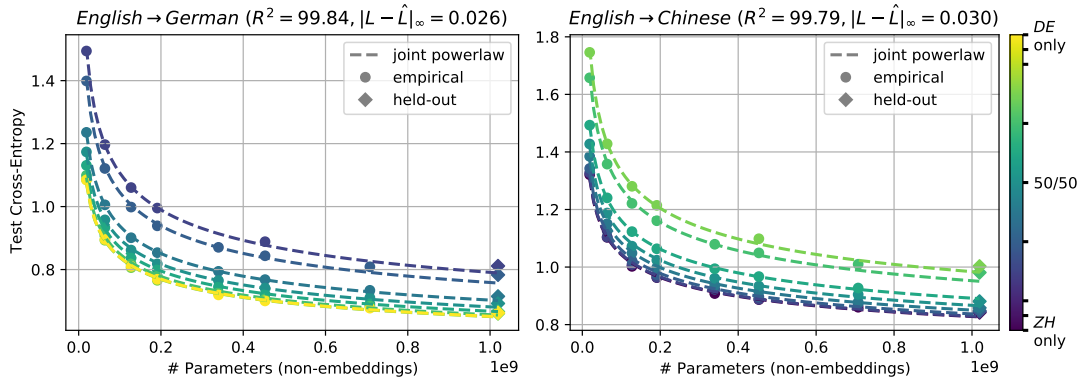

Figure 5. The **joint** scaling law of Equation 7 closely captures the scaling behavior of En→{De, Zh} models. Test loss here is evaluated on in-domain test sets. See Appendix C for similar observations on En→{De, Fr} and {De, Zh}→En models.

doubling the capacity of a multilingual model will reduce its loss by the same $\frac{1}{2^\alpha}$ factor, no matter how the training mixture looks like. This also means that single-language-pair scaling laws can be used to gauge the benefits of scaling multilingual models.

Jointly Modeling Multilingual Scaling Based on the findings above, we make the assumption that the scaling exponents and the irreducible losses are independent of the language pair weights, and propose a **joint** scaling law of the form

$$\mathcal{L}_i(N; p) \approx \beta_{p,i} N^{-\alpha_i} + L_\infty^{(i)}. \tag{7}$$

Figure 5 shows the fit of this joint scaling law for En→{De, Zh} models evaluated on the in-domain test sets. Note that here, we fit a total of 10 parameters for each language pair – 8 for $\beta_{p,i}$'s and two for α_i and $L_\infty^{(i)}$. In contrast, in Figure 2, we used 24 overall parameters to capture the scaling behavior for each language pair. Despite this significant decrease in the number of total fitted parameters, we observe that our joint laws are able to almost completely capture the scaling behavior. We observe a similar phenomenon for out-of-domain test sets and other language pairs (see Appendix C), further suggesting that the joint law accurately describes the scaling behavior of MNMT models.

3.3. Effective Network Capacity for Multilingual Models

We leverage our joint scaling law to examine how MNMT models split their capacity in between the different language pairs. We start by defining the notion of **the effective number of parameters**:

Definition. Consider a multilingual model in which a language pair i has been trained with weight p. We define the effective number of parameters allocated to i, $N_{\text{eff}}^{(i,p)}$, to be equal to the number of parameters necessary for a single-language-pair model solely trained on i to reach the same (test loss) performance as the multilingual model.

Mathematically, $N_{\text{eff}}^{(i,p)}$ can be written as the solution of the

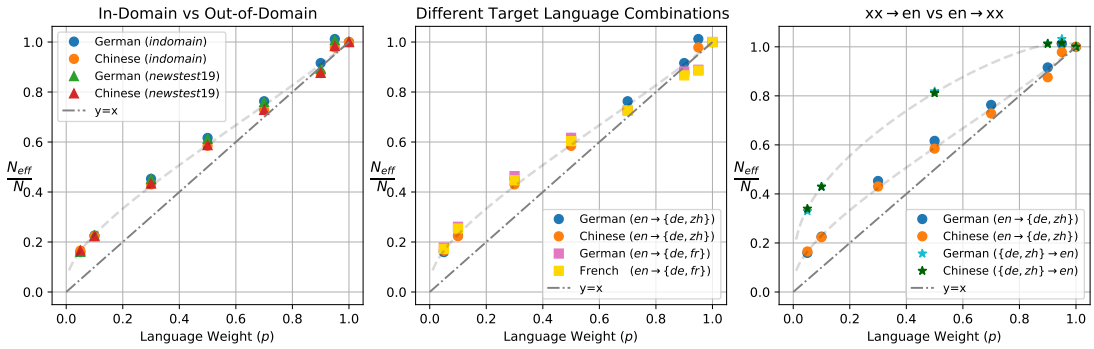

Figure 6. The effective fraction of parameters allocated to each language pair as estimated by our joint scaling laws. Gray dashed lines correspond to the fitted $\hat{f}_i$ described in Equation 12. *Left:* Comparison of the capacity splitting behavior of En→{De, Zh} models for in-domain and out-of-domain test sets. We observe minimal differences between the two. *Center:* Comparison of the capacity splitting behavior for En→{De, Zh} and En→{De, Fr} models. We don't observe any changes in the interaction between the language pairs based on language similarity. *Right:* Comparison of the capacity splitting behavior for translation to and from English. XX→En models exhibit more synergy among the language pairs.

equation

$$\mathcal{L}_i(N; p) = \mathcal{L}_i(N_{\text{eff}}^{(i,p)}; 1). \tag{8}$$

A simple derivation yields that [6]

$$N_{\text{eff}}^{(i,p)} = \left(\frac{\beta_{1,i}}{\beta_{p,i}}\right)^{\frac{1}{\alpha_i}} N. \tag{9}$$

Crucially, our calculations suggest that the fraction of parameters allocated to language pair i, which we denote by $f_i(p)$, is independent of the model size:

$$f_i(p) \equiv N_{\text{eff}}^{(i,p)}/N = \left(\frac{\beta_{1,i}}{\beta_{p,i}}\right)^{\frac{1}{\alpha_i}}. \tag{10}$$

This observation yields a fundamental, scale-independent quantity that can be leveraged for understanding the interactions between the different language pairs in the model.

Figure 6 shows the empirically estimated effective parameter ratios for our models. Several observations are in order:

Consistency Across Domains: In Figure 6 (left), we compare the capacity splitting behavior of the models on in-domain and out-of-domain (newstest19) test sets. Even though the scaling laws coefficients for in-domain and out-of-domain test sets differ, we observe that the capacity splitting behavior is mostly unchanged with different test sets. These findings hint at some measure of universality across test domains on how MNMT models divide their capacity and share their parameters.

Consistency Across Languages Pairs: In Figure 6 (center), we compare the capacity splitting behavior of En→{De, Zh} and En→{De, Fr} models. The conventional wisdom in the MT literature suggests that the tasks in En→{De, Fr} should exhibit a more positive interaction with each other compared to En→{De, Zh}. This is often justified by the intuition that representations are more aligned in related languages and more aligned representations will encourage parameter sharing (Dabre et al., 2017). Surprisingly, our results suggest that the interaction dynamics in En→{De, Fr} and En→{De, Zh} models are not significantly different. In both settings, we observe a relatively neutral multilingual behavior – the performance of an MNMT model of size N trained on language pair i with (sampling) weight p is essentially similar to a single-language-pair model of size pN. In other words, there is minimal synergy among the languages in both setups.

En→XX vs XX→En: In Figure 6 (right), we compare the interaction between the language pairs when translating out of English vs when translating into English.

In stark contrast to the En→XX setting, when translating into English, we observe significant positive synergy among the language pairs. This observation aligns well with recent results in the literature showing multilingual models achieving SOTA performance for translation to English (Chowdhery et al., 2022; Lepikhin et al., 2020). It is unclear if this synergy arises as a specificity of having English as the target language or because multilingual encoding is intrinsically more amenable to parameter sharing than multilingual decoding. Understanding the exact dynamics giving rise to such positive interaction between the language pairs is an exciting open question.

Benefits for Massive Multilingual Models: While we observed minimal synergy between En→XX languages pairs, and therefor minimal gains in *absolute* effective

[6] See Appendix D for details.

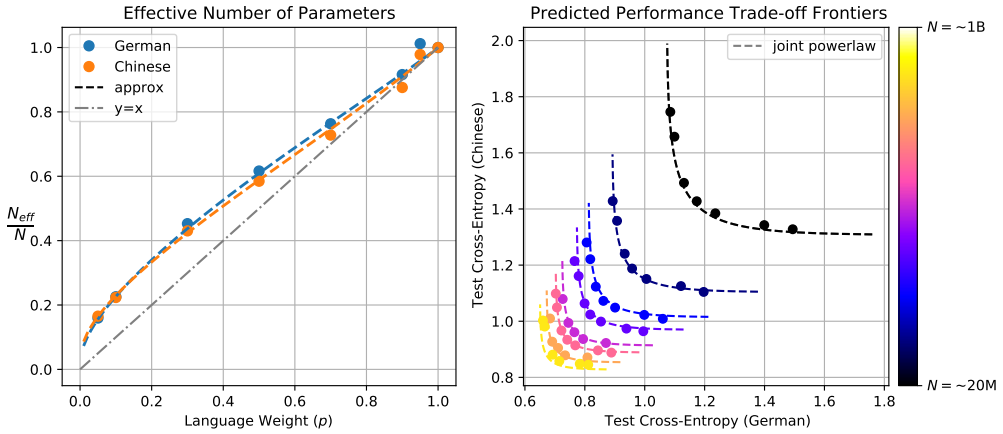

Figure 7. Approximate joint scaling laws described by equations (11) and (12) almost perfectly capture the language pair interactions across all scales. *Left:* The fitted approximation $\hat{f}$ described in Equation 12. *Right:* The predicted performance trade-off frontier (dashed lines) as well as the empirically observed trade-off values.

capacity, if we look at *relative* effective capacity, we can see considerable benefits in using multilingual models for language pairs with small weight. For example, a model trained for En→{De, Zh} with 5% weight on German has an effective capacity of more than 3× a model trained with 5% capacity of this model for only German. These relative gains are even more evident when there is positive task synergy, such as for XX→En, where models train with 5% weight have more than 6× gain in (effective) parameters. This hints that, if these findings generalize beyond the two-task setup[7], then training large multilingual models for training mixtures with a large number of small weight language pairs is significantly more memory efficient than training separate small models for each language pair.

3.4. Guiding Language Balancing

As discussed earlier, one of the areas where multilingual scaling laws can be most impactful is in guiding language balancing/weighting when training large multilingual models, an open problem that has been studied extensively (Arivazhagan et al., 2019a; Wang et al., 2020). However, in its current form, our (joint) scaling law can only be use to decide between weightings that were used for fitting it and cannot be used to predict performance on new, unseen weightings, as $\beta_{p,i}$ needs to be estimated empirically.

To extend to unseen language pair weightings, we instead focus on estimating $f_i(\cdot)$. Given access to $f_i(p)$, accurate prediction of $\mathcal{L}_i(N)$ for **any weighting** can be achieved by using the **single-language-pair scaling law**:

$$\mathcal{L}_i(N;p) = \beta_{1,i}\big(\hat{f}_i(p)N\big)^{-\alpha_i} + L_\infty^{(i)}. \tag{11}$$

As observed in Section 3.3, $f_i(p)$ has a number of desirable properties that makes it easy to estimate: (i) it is invariant to test set and languages, (ii) it is smooth and generally well-behaved. As such, one can achieve an accurate approximation of f with just a few data points.

We utilize this methodology to estimate the full task performance trade-off frontier for En→{De, Zh} models. For estimating $f_i(\cdot)$, we fit an approximate joint scaling law of the form Equation 11, where $\hat{f}_i(\cdot)$ is parameterized as

$$\hat{f}_i(p) = p + c_1 p^{c_2}(1-p)^{c_3} \tag{12}$$

with c_1, c_2, c_3 being fitted coefficients. Figure 7 demonstrates our results; our procedure is able to almost perfectly capture the full task performance frontier across a variety of model scales. With access to such accurate predictions of the performance frontier, a practitioner can precisely determine how to weigh the individual language pairs during training based on her preferences and target model size.

We should note that the choice of function class to fit $f_i(\cdot)$ is highly dependent on the practitioner's computational budget. In our case, we prioritized accuracy and used a flexible function class of the form (12) for fitting. Such flexibility comes with the cost of needing to compute more empirical values to reliably estimate $f(\cdot)$. In the scenarios with more limited computational budget, we have observed that even rudimentary linear approximations of f are able to provide accurate representations of the performance frontier. See Appendix E for examples.

Translation Quality Finally, we note that in the MT literature, quality is often measured via metrics such as BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and BLEURT (Sellam et al., 2020) as opposed to cross-entropy, since the latter doesn't account for the problem of *decoding* translations from the models and is sometimes found to not corre-

[7] see Appendix H for preliminary experiments on models trained on more than two language pairs.

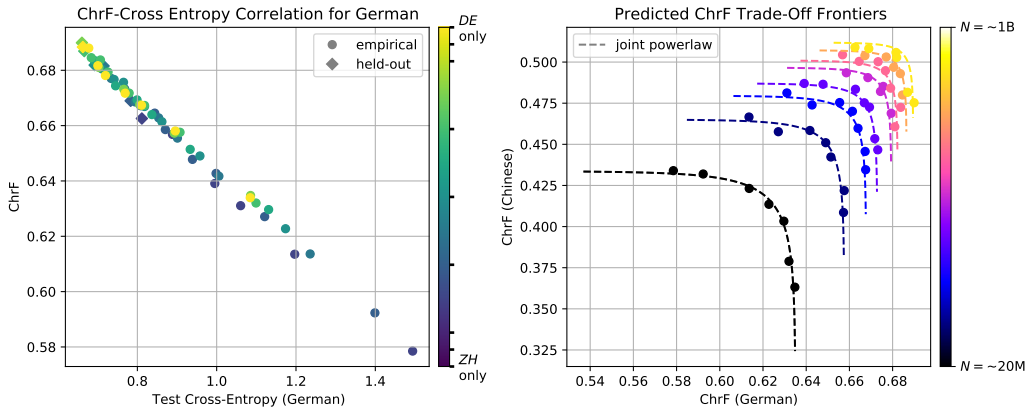

Figure 8. The generation quality behavior of our models as measured by ChrF. *Left:* We observe consistent positive correlations between ChrF and cross-entropy loss. *Right:* Our scaling laws can be used to generate accurate performance trade-off frontiers for ChrF.

late with human preferences (Koehn & Knowles, 2017). As such, MT practitioners might be concerned regarding the applicability of these results for practical applications. To ensure that our findings also apply to the quality of translations, we decode translations from our trained models using beam search (Graves, 2012) and evaluate how their quality changes as we scale the models, using ChrF and BLEURT.

Figure 8 (left) shows cross-entropy and ChrF scores for the En→De language pair of our En→{De, Fr} models, evaluated on the in-domain test set. We find that this automatic metric has an almost-linear relationship with cross-entropy, hinting that our observations also generalize from cross-entropy to generation quality. Figure 8 (right) also shows the predicted ChrF performance trade-off frontier obtained by fitting our joint scaling law (Equation 7) to the ChrF performance on the in-domain test set (parametrizing the effective parameter fraction function as in Equation 12). Our procedure is able to capture this trade-off frontier almost as well as the cross-entropy frontier. Similar findings for the BLEURT metric on out-of-distribution test sets can be found in Appendix F.

4. Conclusions & Future Work

Current state-of-the-art large neural models are moving towards using as much data from as many domains, modalities and languages as possible to unlock exciting new capabilities. Unfortunately, a clear understanding of the behavior of such multitask models at scale is missing. This in turn slows down the model development process since practitioners have to resort to trial and error for balancing their tasks in their models. In this paper, we attempted to take an initial step towards alleviating this problem by performing a large-scale study of the properties of multilingual models.

In particular, we attempted to study this problem from the lens of multilingual machine translation. We showed that, for each language pair and language pair weighting, a power-law describes the evolution of the model test performance as a function of the model size. We examined the dependence of the scaling law parameters on the language pair weights and demonstrated that the scaling exponent and the irreducible loss are independent of the weightings. Using these observations, we provided a novel joint scaling law that succinctly captures the scaling behavior across different model sizes and weightings and used it to define the notion of *effective fraction of parameters* assigned to a language pair ($f_i(\cdot)$). We showed that this quantity robustly captures the language pair interactions and is surprisingly invariant to the similarity of the languages. In the end, we sketched a procedure to use f_i to estimate the task performance trade-off frontier for all model scales.

Future Work In this paper, we studied the scaling behavior of multilingual translation models. Examining whether the conclusions of our work apply to multi-task setups beyond translation is a promising research direction. Most of our conceptual framework and experimental setup can easily be reused for this since there is little difference in the mathematical formulation of the optimization problem and it is likely that similar observations regarding the lack of transfer in data-rich scenarios will be found, as multilinguality can be considered an easier subset of the broader multitask learning challenge.

Furthermore, to keep our investigation tractable, we focused most of our experiments on the two-language-pairs scenario. However, we believe the presented results should be easily extendable to models trained with more languages (see Appendix H). We leave such extensions to future work.

Finally, to simplify the model scaling behavior, we focused our analysis on the data-rich setup. However, in many applications, at least some of the tasks are mid- or low-resource. Extending these results to such scenarios is an interesting future direction.

References

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G. F., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019, 2019a.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G. F., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019, 2019b. URL http://arxiv.org/abs/1907.05019.

Bansal, Y., Ghorbani, B., Garg, A., Zhang, B., Krikun, M., Cherry, C., Neyshabur, B., and Firat, O. Data scaling laws in nmt: The effect of noise and architecture. In *ICML*, 2022.

Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., Breiner, T., Axelrod, V., Riesa, J., Cao, Y., Chen, M. X., Macherey, K., Krikun, M., Wang, P., Gutkin, A., Shah, A., Huang, Y., Chen, Z., Wu, Y., and Hughes, M. Building machine translation systems for the next thousand languages, 2022. URL https://arxiv.org/abs/2205.03983.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Dabre, R., Cromieres, F., and Kurohashi, S. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. In *Proceedings of Machine Translation Summit XVI: Research Track*, pp. 96–107, Nagoya Japan, September 18 – September 22 2017. URL https://aclanthology.org/2017.mtsummit-papers.8.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1723–1732, 2015.

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., García, X., Chelba, C., and Cherry, C. Scaling laws for neural machine translation. *ArXiv*, abs/2109.07740, 2022.

Gordon, M. A., Duh, K., and Kaplan, J. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.478. URL https://aclanthology.org/2021.emnlp-main.478.

Graves, A. Sequence transduction with recurrent neural networks, 2012.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Hernandez, D., Kaplan, J., Henighan, T. J., and McCandlish, S. Scaling laws for transfer. *ArXiv*, abs/2102.01293, 2021.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32, 2019.

Hutter, M. Learning curve theory. *CoRR*, abs/2102.04074, 2021. URL https://arxiv.org/abs/2102.04074.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Koehn, P. and Knowles, R. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204.

Kudo, T. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL https://aclanthology.org/P18-1007.

Kurin, V., De Palma, A., Kostrikov, I., Whiteson, S., and Kumar, M. P. In defense of the unitary scalarization for deep multi-task learning. *arXiv preprint arXiv:2201.04122*, 2022.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. Multi-task sequence to sequence learning, 2015. URL https://arxiv.org/abs/1511.06114.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Popović, M. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2019.

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Sellam, T., Das, D., and Parikh, A. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Shazeer, N. M. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. *ArXiv*, abs/1804.04235, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, W., Watanabe, T., Hughes, M., Nakagawa, T., and Chelba, C. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 133–143, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6314. URL https://aclanthology.org/W18-6314.

Wang, X., Tsvetkov, Y., and Neubig, G. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8526–8537, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.754. URL https://aclanthology.org/2020.acl-main.754.

Wang, Z., Tsvetkov, Y., Firat, O., and Cao, Y. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=F1vEjWK-lH_.

Xin, D., Ghorbani, B., Garg, A., Firat, O., and Gilmer, J. Do current multi-task optimization methods in deep learning even help? *Advances in neural information processing systems*, 2022.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhang, B., Ghorbani, B., Bapna, A., Cheng, Y., Garcia, X., Shen, J., and Firat, O. Examining scaling and transfer of language model architectures for machine translation. *arXiv preprint arXiv:2202.00528*, 2022.

A. Model Sizes and Hyperparameters

Enc. Layers	Dec. Layers	Emb. Dim	# Heads	Head Dim	MLP dim	Vocab Size	# Parameters	Corrected # Parameters
2	2	512	8	64	2048	128k	149,953,024	18,881,024
3	3	768	12	64	3072	128k	260,322,816	63,714,816
6	6	768	12	64	3072	128k	324,035,328	127,427,328
9	9	768	12	64	3072	128k	387,747,840	191,139,840
9	9	1024	16	64	4096	128k	601,931,776	339,787,776
12	12	1024	16	64	4096	128k	715,193,344	453,049,344
12	12	1280	16	80	5120	128k	1,035,876,864	707,869,184
12	12	1536	16	96	6144	128k	1,412,528,128	1,019,312,128

B. Individual Scaling Laws Fits

B.1. Out-of-Domain

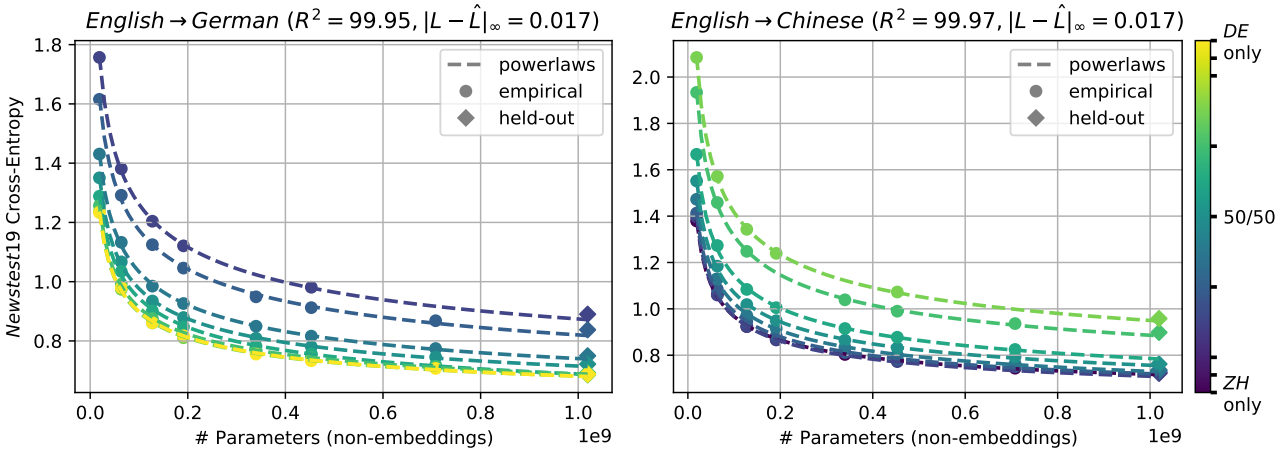

Figure 9. The evolution with model size of the cross-entropy loss on the *newstest19* test set for En→{De, Fr} models, as well as the fitted scaling laws. The color represents the weighting of the languages. Note that we don't show the *zero-shot* behavior.

B.2. English→German, French

B.3. German, Chinese→English

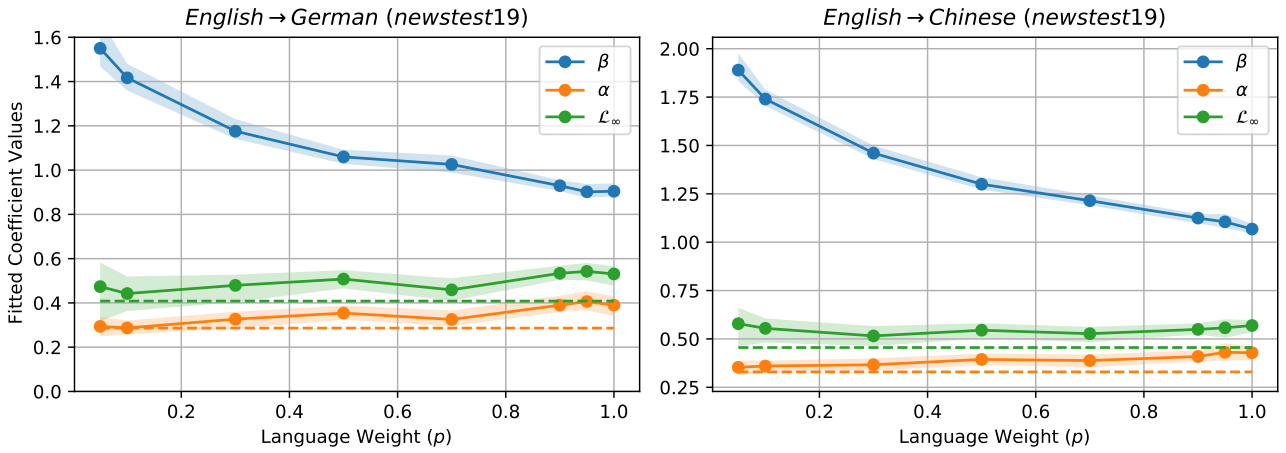

Figure 10. Coefficient values, for scaling laws fitted on *newstest2019*, for German (left) and French (right) as a function of the language weight, with the shaded region representing the standard deviation. The dashed lines represent the value of jointly fitted coefficients from Equation 7

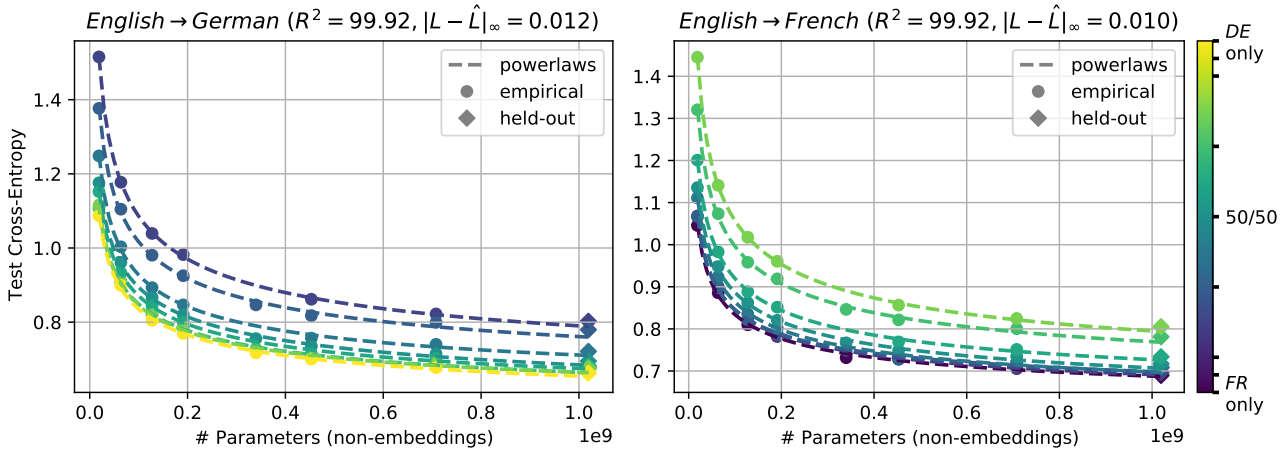

Figure 11. The evolution of the (in-domain) test cross-entropy loss with model size for En→{De, Fr} models, as well as the fitted scaling laws. The color represents the weighting of the languages. Note that we don't show the *zero-shot* behavior.

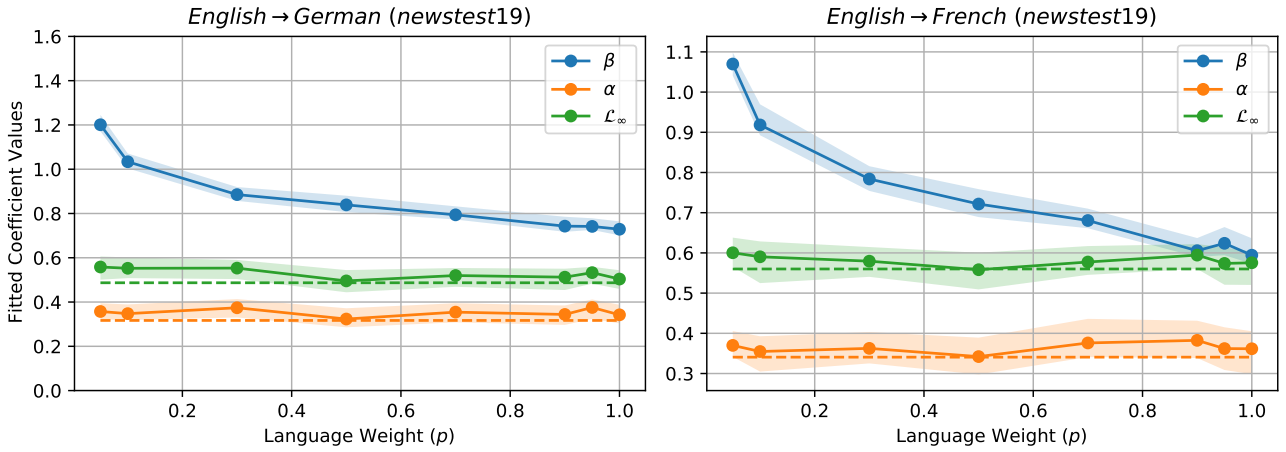

Figure 12. Coefficient values for German (left) and French (right) as a function of the language weight, with the shaded region representing the standard deviation. The dashed lines represent the value of jointly fitted coefficients from Equation 7

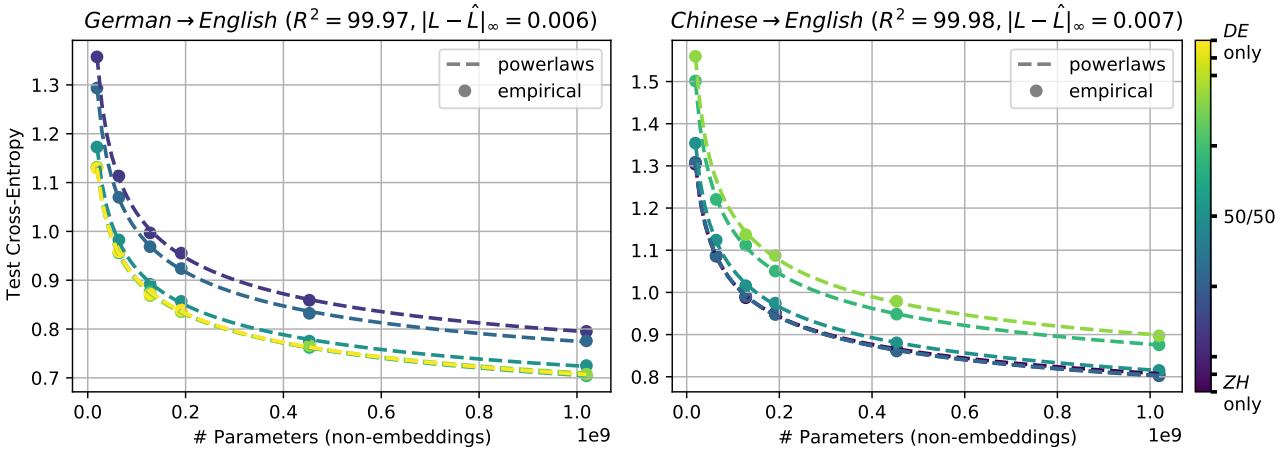

Figure 13. The evolution of the (in-domain) test cross-entropy loss with model size for {De, Zh}→En models, as well as the fitted scaling laws. The color represents the weighting of the languages. Note that we don't show the *zero-shot* behavior.

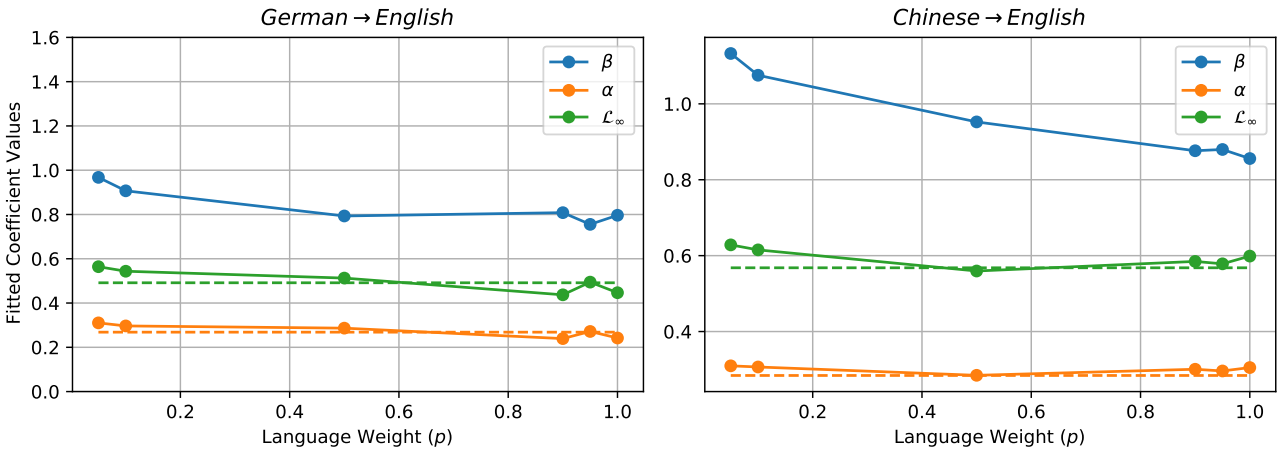

Figure 14. Coefficient values for German (left) and French (right) into English as a function of the language weight. The dashed lines represent the value of jointly fitted coefficients from Equation 7.We omit uncertainty estimates since less model capacities were used to fit the scaling laws, and therefor these estimates would be unreliable.

C. Joint Scaling Law Fits

C.1. Out-of-Domain

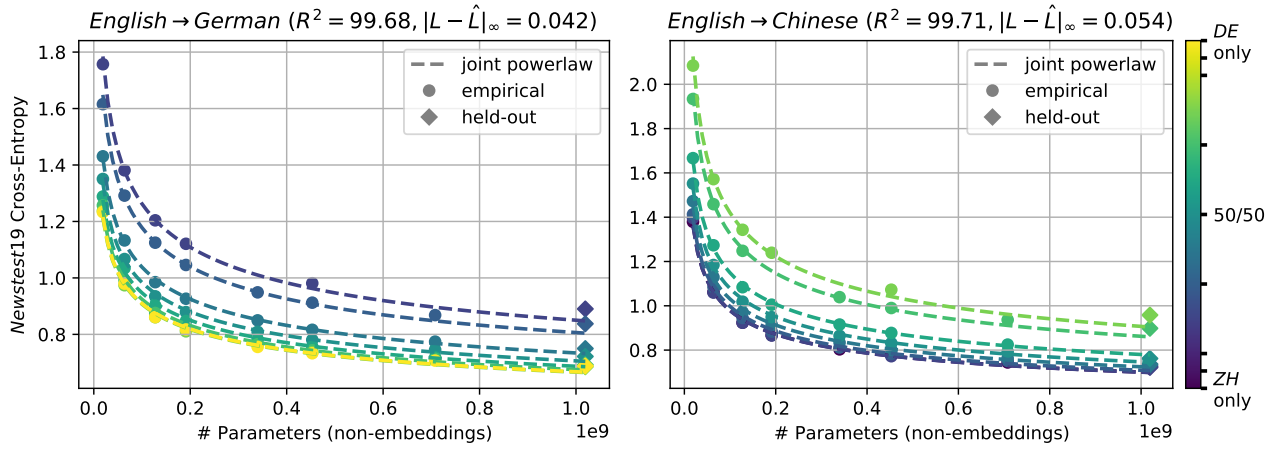

Figure 15. The **joint** scaling law (Equation 7) fitted to models trained for En→{De, Zh} models. Test loss here is evaluated on the *newstest2019* test set.

C.2. English→{German, French}

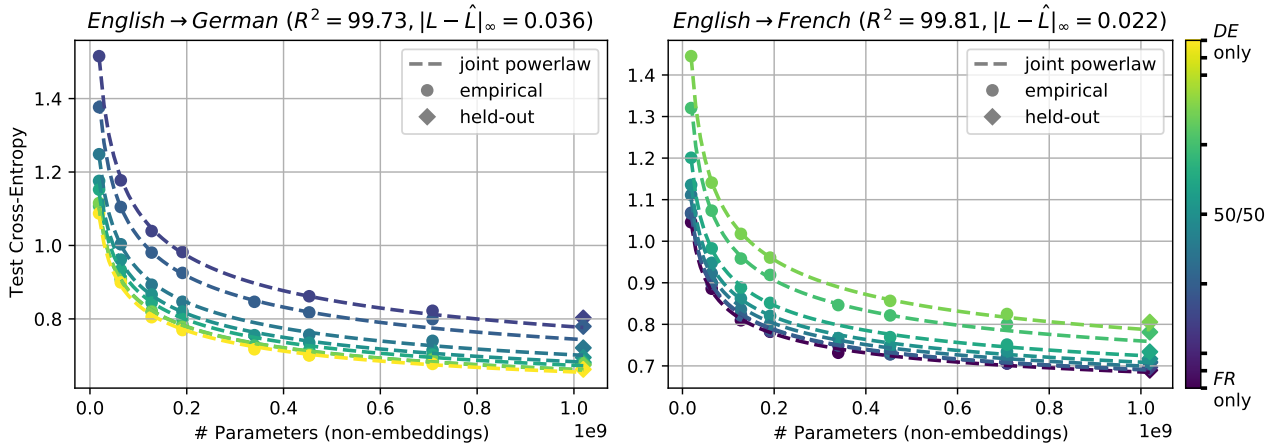

Figure 16. The **joint** scaling law (Equation 7) fitted to models trained for En→{De, Fr} models. Test loss here is evaluated on in-domain test sets.

C.3. {German, Chinese}→English

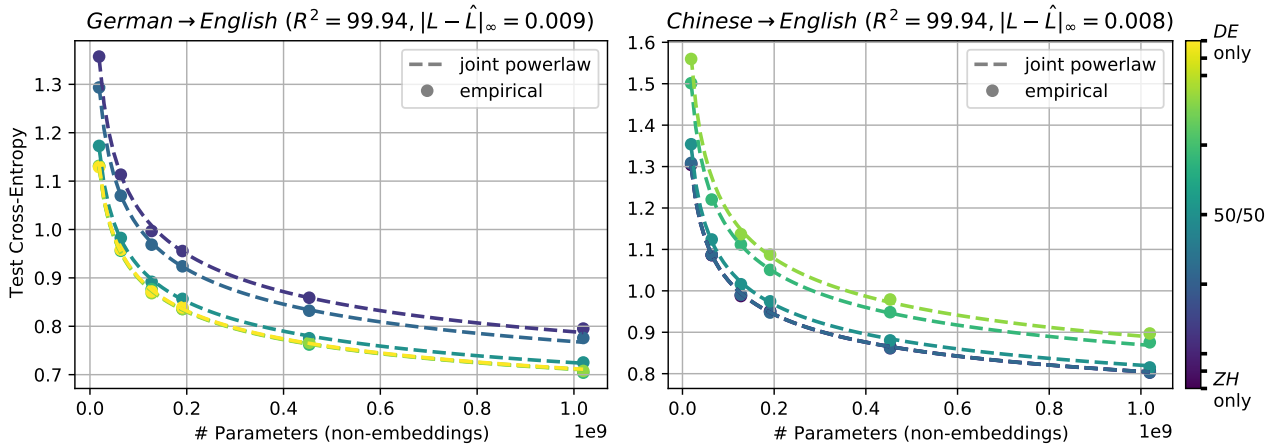

Figure 17. The **joint** scaling law (Equation 7) fitted to models trained for {De, Zh}→En models. Test loss here is evaluated on in-domain test sets.

D. Derivation of the Effective Number of Parameters

$$
\begin{aligned}
\mathcal{L}_i(N;p) &= \beta_{p,i} N^{-\alpha_i} + L_\infty^{(i)} \\
&= \beta_{1,i} \left(\frac{\beta_{p,i}}{\beta_{1,i}}\right) N^{-\alpha_i} + L_\infty^{(i)} \\
&= \beta_{1,i} \left(\left(\frac{\beta_{p,i}}{\beta_{1,i}}\right)^{-\frac{1}{\alpha_i}} N\right)^{-\alpha_i} + L_\infty^{(i)} \\
&= \beta_{1,i} \left(\left(\frac{\beta_{1,i}}{\beta_{p,i}}\right)^{\frac{1}{\alpha_i}} N\right)^{-\alpha_i} + L_\infty^{(i)} \\
&= \beta_{1,i} N_{\text{eff}}^{-\alpha_i} + L_\infty^{(i)} \\
&= \mathcal{L}_i(N_{\text{eff}};p)
\end{aligned}
$$

E. Other Approximations to the Effective Parameter Ratio

We use a linear approximation of the form

$$\hat{f}_i(p) = c_1(p-1) + 1. \tag{13}$$

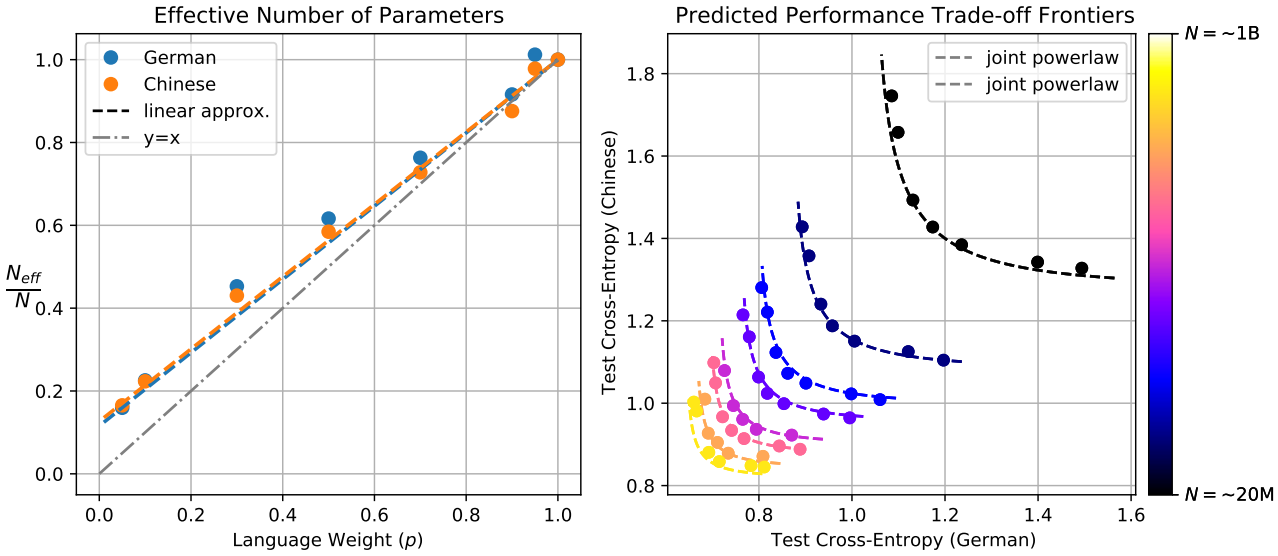

Figure 18. Approximate joint scaling laws described by equations (11) and (13) is able to capture the task interactions across all scales well, even with single fitted coefficient for ratio function. *Left:* The fitted approximation $\hat{f}$ described in Equation 12. *Right:* The predicted performance trade-off frontier (dashed lines) as well as the empirically observed trade-off values.

F. Translation Quality

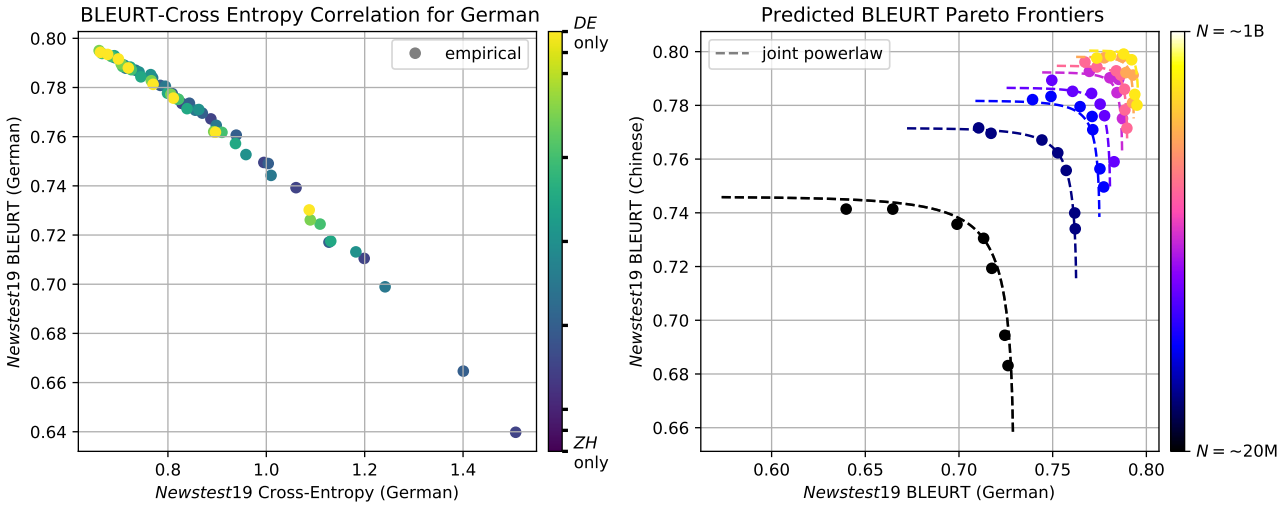

Figure 19. (left) shows cross-entropy and BLEURT scores for the En→De language pair of our En→{De, Fr} models, evaluated on the *newstest19* test set. We find that this automatic metric has an almost-linear relationship with cross-entropy, hinting that our observations also generalize from cross-entropy to generation quality. Figure 8 (right) also shows the predicted BLEURT performance trade-off frontier obtained by fitting our joint scaling law (Equation 7) to the BLEURT performance on the *newstest19* test set (parametrizing the effective parameter fraction function as in Equation 12).

G. Convergence Correction

Due to *implicit* scalarization, models trained with very little task weight (< 0.1) will see less than a full epoch of that task's data, even when trained with 1M steps. I our experiments we saw that this was causing problems in the fit the scaling laws due to an *undertraining* of our largest models.

To mitigate this problem without training these models for a prohibitively large number of steps, we apply recent findings in learning curve (Hutter, 2021) to estimate the performance of largest models trained with $p \leq 0.05$ task weight at convergence, by fitting a power-law to the performance evolution as training progresses, and predicting the performance of these models at 2.5M steps. This only affect two models per scenario considered.

H. Extension to more than two languages/tasks

As an early effort to understand if our findings apply to more than two tasks, we trained various model sizes for to translate into **three** languages (EEn→{De, Fr, Zh}), and compared the predictions using the scaling laws for models trained on two language pairs (En→{De, Zh} and En→{De, Fr}).

Figure Figure 20 shows the results. Overall we find that (combination of) the joint scaling laws fitted on models trained on two language pairs predict well the performance of models trained for three language pairs, showing that the invariances found in previous sections generalize to settings with more than two tasks. These results also hint that computation of effective parameters counts for multi-task models with many tasks can be simplified and made more tractable by training models with much smaller subset of tasks.

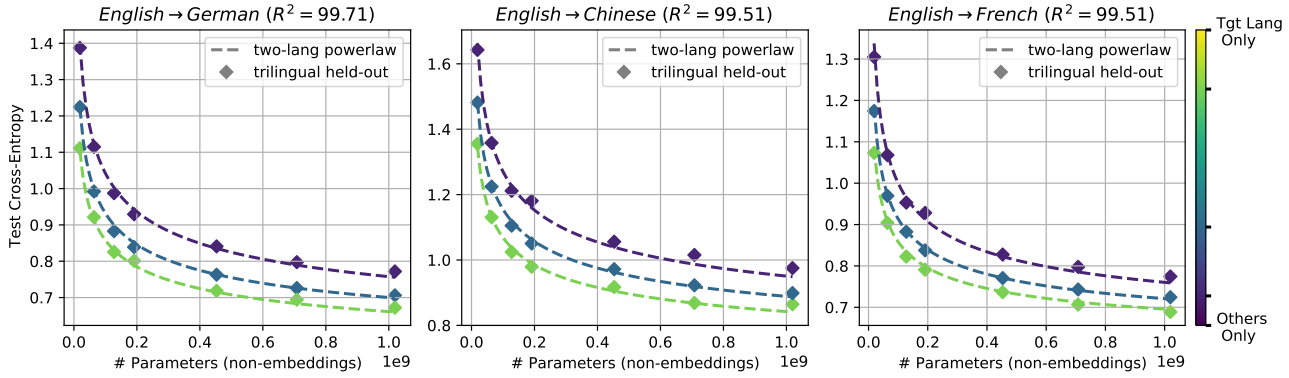

Figure 20. The evolution of the (in-domain) test cross-entropy loss with model size for En→{De, Fr, Zh} models, as well as the fitted scaling laws fitted for En→{De, Zh} (left and middle) and En→{De, Fr} (right). The color represents the weighting of the languages. Note that we don't show the *zero-shot* behavior.