SCALING LAWS FOR DOWNSTREAM TASK PERFORMANCE OF LARGE LANGUAGE MODELS

Berivan Isik*
Stanford University
berivan.isik@stanford.edu

Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii
Google Research

Sanmi Koyejo
Stanford University & Google DeepMind

ABSTRACT

Scaling laws provide important insights that can guide the design of large language models (LLMs). Existing work has primarily focused on studying scaling laws for pretraining (upstream) loss. However, in transfer learning settings, in which LLMs are pretrained on an unsupervised dataset and then finetuned on a downstream task, we often also care about the downstream performance. In this work, we study the scaling behavior in a transfer learning setting, where LLMs are finetuned for machine translation tasks. Specifically, we investigate how the choice of the *pretraining* data and its size affect downstream performance (translation quality) as judged by two metrics: downstream cross-entropy and BLEU score. Our experiments indicate that the size of the finetuning dataset and the distribution alignment between the pretraining and downstream data significantly influence the scaling behavior. With sufficient alignment, both downstream cross-entropy and BLEU score improve monotonically with more pretraining data. In such cases, we show that it is possible to predict the downstream BLEU score with good accuracy using a log-law. However, there are also cases where moderate misalignment causes the BLEU score to fluctuate or get worse with more pretraining, whereas downstream cross-entropy monotonically improves. By analyzing these observations, we provide new practical insights for choosing appropriate pretraining data.

1 INTRODUCTION

There has been extensive research on scaling laws for *upstream* perplexity or cross-entropy loss (i.e., evaluated on pretraining data) of large language models (LLMs), demonstrating that these quantities can be well predicted using power laws (Kaplan et al., 2020; Hoffmann et al., 2022; Gordon et al., 2021; Hernandez et al., 2022; Fernandes et al., 2023; Bansal et al., 2022; Henighan et al., 2020; Johnson et al., 2018). However, in practical applications, LLMs often undergo transfer learning–they are first pretrained on unsupervised data and then finetuned for specific downstream[1] tasks such as coding or translation. The question of whether scaling laws can be used to predict downstream task performance is critical, yet remains largely unanswered Hernandez et al. (2021); Tay et al. (2021). Here, the term *task performance* refers to metrics that measure task-related quantities such as accuracy and BLEU score, which are different from next-token prediction metrics such as cross-entropy.

In this work, we study scaling laws for transfer learning and focus on machine translation tasks. Specifically, we look into the relation between the pretraining dataset size and the *downstream task performance* after finetuning on the task. We find that, in addition to the finetuning data size and the

*Work done while interning at Google Research.

[1]We use the term *downstream* to refer to the finetuning task or metrics computed on it, and the term *upstream* to refer to the metrics computed on the pretraining dataset.

choice of the performance metric, this relation fundamentally depends on the alignment between the pretraining data and the downstream task. While similar observations have been made in different contexts in the transfer learning literature (Tamkin et al., 2020; Agostinelli et al., 2022), our work provides new insights and concrete scaling laws for the downstream performance of LLMs.

We carry out systematic experiments in which we pretrain LLMs on multilingual unsupervised datasets and then finetune them on several machine translation tasks. Across the experiments, we vary the type of pretraining data (to control the degree of distribution alignment with the downstream task) and the finetuning data size. We study two metrics: *downstream* BLEU Papineni et al. (2002) score[2] and *downstream* cross-entropy. We find that in settings where the distributions are well-aligned, both BLEU and *downstream* cross-entropy improve monotonically with more pretraining. In these settings, we demonstrate that the BLEU score can be well predicted using the following log-law: $f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}$, where D_p denotes the size of the pretraining data, and A, α, β are the coefficients to be fit. We further propose a power-law $L(D_p) = E + \frac{A}{D_p^{\alpha}}$ for the *downstream* cross-entropy as the pretraining data scales – echoing similar laws developed for the *upstream* cross-entropy as a function of the pretraining dataset size (Kaplan et al., 2020; Hoffmann et al., 2022) and *downstream* cross-entropy as a function of the finetuning dataset size (Hernandez et al., 2021).

However, when distributions are not sufficiently aligned and the finetuning data size is relatively small, we find that there are cases where the BLEU score exhibits an unclear, non-monotonic behavior, whereas the *downstream* cross-entropy still improves monotonically following a power-law. This observation suggests that using cross-entropy as a proxy for task-related metrics like BLEU score may lead to critical misjudgments in practice if used to make decisions about the "relevance" of the pretraining data for the downstream task or the required size of the pretraining data for the target downstream performance. Arguing that the value of pretraining data should be evaluated using *downstream task-related metrics like BLEU score*, we propose a practical guide for such an assessment by leveraging the proposed scaling law for BLEU score in Appendix C.

Finally, our empirical studies suggest that pretraining brings little to no improvement on the BLEU score when the finetuning (translation) dataset is already large enough, complementing the findings of Hernandez et al. (2021). We discuss related work in Appendix B

2 Scaling Laws for Transfer Learning

2.1 A Scaling Law for the BLEU Score

Different from cross-entropy and perplexity, which follow a power-law scaling behavior Kaplan et al. (2020); Hoffmann et al. (2022), we find out that BLEU score scales closer to a log-law, as evident from Figures 1, 2, and 5. Therefore, we propose the following scaling law for BLEU score as a function of the pretraining dataset size D_p:

$$f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}, \tag{1}$$

where A, α, and β are coefficients to be fit. We notice that these coefficients depend on how aligned the pretraining dataset with the target downstream task (translation from language 1 to language 2) and how large the finetuning (translation) dataset is. With extensive experiments across several translation tasks and multilingual pretrained models, we demonstrate that the law in (1) indeed well describes BLEU score scaling, with a small prediction error which we quantify in Appendix E.2.

2.2 Is Cross-Entropy Loss Always a Good Metric?

We also compare the *downstream* cross-entropy and the BLEU score empirically as prior work has made the assumption that *upstream* or *downstream* cross-entropy is a good indicator for a model's *downstream task performance*. Following the well-understood scaling behavior of the *upstream* cross-entropy as a function of the pretraining dataset size Kaplan et al. (2020); Hoffmann et al. (2022), we demonstrate that the same scaling law can also describe the *downstream* cross-entropy loss as

$$L(D_p) = E + \frac{A}{D_p^{\alpha}}, \tag{2}$$

[2] In the rest of the paper, we will drop "downstream" when we refer to the downstream BLEU score.

where E, A, and α are the coefficients to be optimized. Throughout the paper, we report BLEU score and cross-entropy together for a direct comparison and discover several cases where the two metrics do not correlate well. This supports some of the findings of Ghorbani et al. (2021) suggesting inconsistency between the BLEU score and the cross-entropy, but also shows that the exponential relationship (between the two metrics) advocated by Gordon et al. (2021) does not always hold. More specifically, our empirical results show that while cross-entropy loss always monotonically decreases (with appropriate learning rate) as the pretraining dataset size increases, BLEU score may show a non-monotonic trend when the pretraining data is not sufficiently aligned with the task. For instance, in Figure 5-(*top, right*), increasing the en-MC4, de-MC4, or ro-MC4 pretraining datasets' size sometimes decreases the BLEU score on WMT-15 English-to-French (en-fr) translation task. Even though they may initially follow the law in (1) for smaller pretraining dataset sizes, the scaling law breaks for larger data for these datasets and task. Overall, the BLEU score never reaches a good value compared to other pretraining datasets that include some amount of French – indicating that pretraining datasets that do not include French are not aligned enough with this particular translation task. However, if we were to look at only the cross-entropy loss in Figure 5-(*bottom, right*), we would conclude that all the pretraining datasets bring noticeable improvements to the model and they all are worth adding into the pretraining data – which would be a poor decision.

2.3 When Do Scaling Laws Fall Short in Transfer Learning?

While the cross-entropy loss always follows a monotonically decreasing trend which can be captured by the scaling law in (2), we do not always see a monotonic increase in the BLEU score when increasing the pretraining dataset size (see Figure 2-(*top, center*) and Figure 5-(*top, right*)). We observe that this only happens when the pretraining dataset is not sufficiently aligned with the translation task – which results in low BLEU scores overall compared to models that were pretrained in other datasets. For the pretrained models that lead to high BLEU scores after finetuning, we consistently see that the BLEU score increases monotonically and can be well described with the scaling law in (1). Therefore, whether the scaling law could fit the empirical BLEU scores or not could be a good first-check in assessing the value of pretraining data for the downstream (translation) task. We elaborate more on this in Appendix C and propose a guide for assessing the value of pretraining dataset for a target downstream task.

3 Experimental Setup

We first pretrain a model without doing more than one pass over any of the examples. Then, we finetune selected checkpoints of the pretrained model. Naturally, there is a one-to-one mapping between the checkpoint number and the number of pretraining tokens seen. This way, we collect pairs of (number of pretraining tokens, BLEU score) and (number of pretraining tokens, *downstream* cross-entropy loss) to analyze them with the proposed scaling laws in (1) and (2). All the plots are on a $\log$-$\log$ scale. See Appendix D for details on how to optimize the coefficients of the scaling laws.

We use the 3-billion and 770-million encoder-decoder T5 models, same as the T5-3B and T5-Large model in Abnar et al. (2022). More details about model architecture and hyperparameters are provided in Appendix D. We use the English (en), German (de), French (fr), and Romanian (ro) portions of the MC4 dataset. We experiment with both pretraining on these languages individually as well as mixing pairs of languages. In Figure 1, we present results for the models pretrained on (*left*) a mixture of 50% en-MC4 + 50% de-MC4, (*center*) a mixture of 50% en-MC4 + 50% fr-MC4, and (*right*) a mixture of 50% en-MC4 + 50% ro-MC4 – meaning that 50% of one pretraining batch is sampled from en-MC4 and the other 50% is sampled from the other language. In Figure 2, we show results for the models pretrained only on en-MC4. In Appendix E, in addition to these, we also present results for the models pretrained on a mixture of 30% en-MC4 + 70%-fr and a mixture of 70% en-MC4 + 30%-fr as well as models pretrained only on de-MC4, only on fr-MC4, and only on ro-MC4. We finetune the pretrained models on WMT-17 en-de (Bojar et al., 2017), WMT-15 en-fr (Bojar et al., 2014), and WMT-16 en-ro (Bojar et al., 2016), separately. To understand the effect of the finetuning dataset size on the scaling laws, we sometimes use a smaller randomly sampled portion from these translation datasets and indicate the number of tokens used.

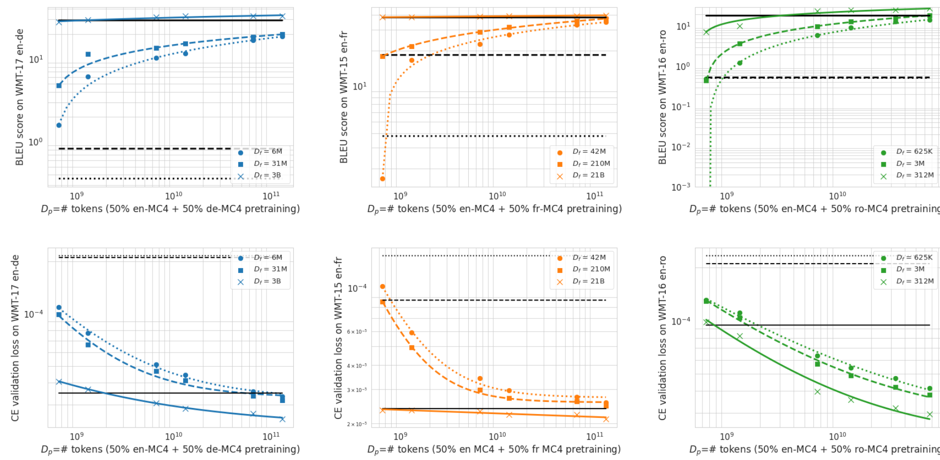

Figure 1: **(top) BLEU score vs pretraining dataset size: $\mathbf{f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}}$.** ***(left)*** WMT-17 en-de translation task. Pretraining dataset has 50% en-MC4 + 50% de-MC4. ***(center)*** WMT-15 en-fr translation task. Pretraining dataset has 50% en-MC4 and 50% fr-MC4. ***(right)*** WMT-16 en-ro translation task. Pretraining dataset has 50% en-MC4 + 50% ro-MC4. **(bottom) Cross-entropy (CE) validation loss vs pretraining dataset size: $\mathbf{L(D_p) = E + \frac{A}{D_p^{\alpha}}}$.** Same models as the top row. The markers are the actual experimental results and the black horizontal curves correspond to the non-pretrained model directly trained on the task dataset. **The finetuning dataset size D_f increases in the order of dotted-dashed-solid for all the curves including the black horizontal lines.**

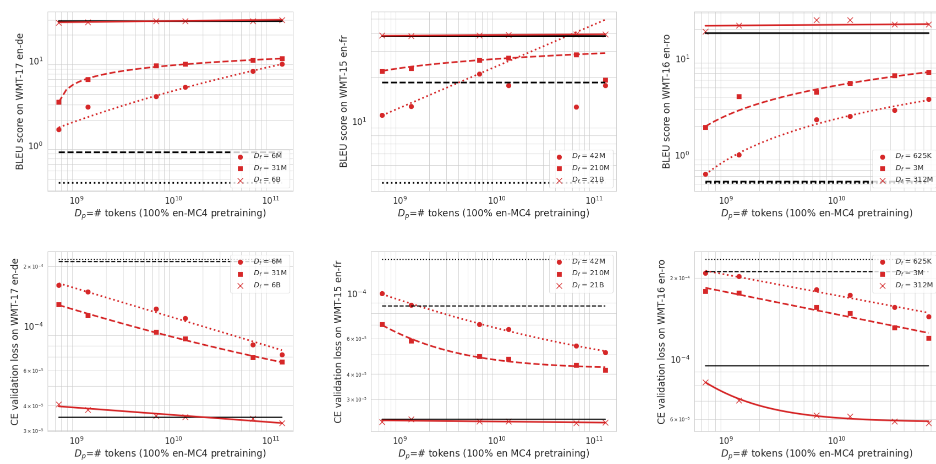

Figure 2: Same as Figure 1 but the pretraining dataset is 100% en-MC4 in all plots.

4 Results and Analysis

In Figure 1, we analyze the models that are pretrained on different portions of *(left)* a mixture of 50% en-MC4 + 50% de-MC4, *(center)* a mixture of 50% en-MC4 + 50% fr-MC4, and *(right)* a mixture of 50% en-MC4 + 50% ro-MC4. These models are then finetuned on different portions of *(left)* en-de, *(center)* en-fr, and *(right)* en-ro translation datasets. In the top row, we report the BLEU score and, in the bottom row, we report the *downstream* cross-entropy loss. The dotted, dashed, and solid lines correspond to the scaling laws in (1) and (2) for different finetuning dataset sizes D_f. The black lines correspond to "non-pretrained" models that are directly trained on different portions of the finetuning dataset. In all cases, the scaling laws fit well to the empirical results (the markers). As expected, as the finetuning dataset size increases (going in the order of dotted-dashed-solid lines), the BLEU score increases and the cross-entropy loss decreases smoothly and monotonically. Similarly, as the pretraining dataset size D_p increases (along the x-axis), we see improvements in both metrics. Notice that the improvements by an increase in the pretraining dataset size is more effective for smaller finetuning datasets. When the finetuning dataset is large enough (e.g., solid lines), BLEU score is more or less constant regardless of the pretraining dataset size. In fact, we see little to no improvement of pretraining compared to the non-pretrained models (black lines) when the finetuning dataset is large. In Figure 2, we change the pretraining dataset to 100% en-MC4 in all plots. Intuitively, we expect this dataset to be less aligned with the translation tasks than the multilingual pairs in Figure 1 since it does not include one of the languages in the translation tasks. Indeed, we see smaller BLEU score and higher cross-entropy loss in general for the same finetuning dataset size. Most of the conclusions from Figure 1 carry over to the results in Figure 2. One noticeable difference is in the BLEU scores for the en-fr translation task (*center*). We see that, for $D_f = 42M$ and $D_f = 210M$, the scaling law for BLEU score actually breaks once the pretraining dataset size passes a threshold while the cross-entropy loss scales as expected. This is counter-intuitive because the BLEU score sometimes decreases for larger pretraining dataset. Notice that this break in scaling law does not happen in en-de or en-ro translation tasks as the scaling law fits well to the pretraining for these tasks. We investigate this mismatch between the BLEU score and the *downstream* cross-entropy when the alignment is not sufficient in detail in Appendix E.

5 Discussion and Conclusion

We study the scaling behavior of the downstream performance of LLMs as the pretraining data grows and propose scaling laws for both *downstream* cross-entropy and the BLEU score. We demonstrate that the scaling behavior is significantly influenced by (1) the degree of alignment between the pretraining and the downstream data and (2) the finetuning dataset size. In favorable cases where the distributions are sufficiently aligned, we show that BLEU score can be accurately predicted using a log scaling law. However, with less alignment, there are cases where BLEU score fluctuates unpredictably whereas *downstream* cross-entropy improves monotonically. We also observe that when the finetuning dataset size is sufficiently large, pretraining has little to no value. Our findings highlight the importance of studying downstream performance metrics and not making decisions solely based on cross-entropy (whether upstream or downstream). This echoes the findings of Schaeffer et al. (2023) about the discrepancy in behavior between smooth and non-smooth metrics when models are scaled. **We refer the reader to the Appendix for the full details of our work.**

6 Broader Impact

In this work, we study the scaling behavior of downstream task performance as a function of the pretraining data size and analyze several factors that affect this behavior. We believe our findings are important to avoid costly pretraining on irrelevant datasets – reducing the carbon footprint of training LLMs and freeing compute resources for more impactful tasks.

References

Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=V3C8p78sDa`.

Andrea Agostinelli, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. Transferability metrics for selecting source model ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7936–7946, 2022.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pp. 1466–1482. PMLR, 2022.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, pp. 2309–2313. IEEE, 2019.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pp. 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W16/W16-2301`.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W17-4717`.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 12–58, 2014.

Cheng-Han Chiang and Hung-yi Lee. On the transferability of pre-trained language models: A study from artificial datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10518–10525, 2022.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. Using similarity measures to select pretraining data for ner. 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)*, pp. 1460–1470, 2019.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling laws for multilingual neural machine translation. *arXiv preprint arXiv:2302.09650*, 2023.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. In *International Conference on Learning Representations*, 2021.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *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`.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *International Conference on Machine Learning*, pp. 9201–9225. PMLR, 2022.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

Shibal Ibrahim, Natalia Ponomareva, and Rahul Mazumder. Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 693–709. Springer, 2022.

Achin Jain, Gurumurthy Swaminathan, Paolo Favaro, Hao Yang, Avinash Ravichandran, Hrayr Harutyunyan, Alessandro Achille, Onkar Dabeer, Bernt Schiele, Ashwin Swaminathan, et al. A meta-learning approach to predicting performance and data requirements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3623–3632, 2023.

Mark Johnson, Peter Anderson, Mark Dras, and Mark Steedman. Predicting accuracy on large datasets from smaller pilot data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 450–455, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Taku Kudo. 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, 2018.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP 2018*, pp. 66, 2018.

Hiroaki Mikami, Kenji Fukumizu, Shogo Murai, Shuji Suzuki, Yuta Kikuchi, Taiji Suzuki, Shin-ichi Maeda, and Kohei Hayashi. A scaling law for syn2real transfer: How much is your pre-training effective? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 477–492. Springer, 2022.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=j5BuTrEj35`.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pp. 7294–7305. PMLR, 2020.

Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

Barbara Plank and Gertjan Van Noord. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1566–1576, 2011.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=ITw9edRDlD`.

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

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Object detection from scratch with deep supervision. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):398–412, 2019.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. Investigating transferability in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1393–1401, 2020.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations*, 2021.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S Morcos. D4: Improving llm pretraining via document de-duplication and diversification. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1395–1405, 2019.

Vincent Van Asch and Walter Daelemans. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pp. 31–36, 2010.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pp. 12133–12143. PMLR, 2021.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Biao Zhang, Behrooz Ghorbani, Ankur Bapna, Yong Cheng, Xavier Garcia, Jonathan Shen, and Orhan Firat. Examining scaling and transfer of language model architectures for machine translation. In *International Conference on Machine Learning*, pp. 26176–26192. PMLR, 2022.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. Scaling law for document neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8290–8303, 2023.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33: 3833–3845, 2020.

A Contributions

Our contributions and main findings can be summarized as follows.

- We carry out systematic experiments on 770-million and 3-billion encoder-decoder T5 Raffel et al. (2020) models to study how downstream performance, measured by *downstream* cross-entropy and BLEU score, scales with the pretraining dataset size. For pretraining, we experiment with different subsets of the Multilingual C4 (MC4) dataset (Raffel et al., 2020), including English (en), German (de), French (fr), and Romanian (ro). For finetuning, we study the following translation tasks: WMT-17 en-de Bojar et al. (2017), WMT-15 en-fr Bojar et al. (2014), and WMT-16 en-ro Bojar et al. (2016).
- We observe that, when the distributions of the pretraining and downstream tasks are well-aligned, the BLEU score and *downstream* cross-entropy improve monotonically with more pretraining. For BLEU score, we propose a new log scaling law and show that it has good predictive accuracy. item When the distributions are not sufficiently aligned and the finetuning data size is relatively small, the BLEU score fluctuates or even gets worse with more pretraining–losing the monotonic scaling behavior. In these same settings, we find that the *downstream* cross-entropy still scales monotonically according to a power-law.
- We argue that the value of pretraining data should be evaluated using *downstream task-related metrics like BLEU score* and propose a practical guide for such an assessment by leveraging the proposed scaling law for BLEU score.

B Related Work

Scaling laws for transformers. Scaling laws for LLMs have attracted significant attention as they can inform the decisions about key design choices such as model size and the type and size of the pretraining data Kaplan et al. (2020); Hoffmann et al. (2022); Hernandez et al. (2021). Most of the pioneering work has focused on how *upstream* cross-entropy loss or perplexity scales with more pretraining data, larger models, or longer training Kaplan et al. (2020); Hoffmann et al. (2022). Follow-up works have analyzed scaling behavior of translation models (Ghorbani et al., 2021; Zhuocheng et al., 2023; Gordon et al., 2021; Fernandes et al., 2023; Bansal et al., 2022; Zhang et al., 2022), studied theoretical foundation behind scaling laws Sharma & Kaplan (2020); Hutter (2021); Bahri et al. (2021), or extended the laws to the vision models Zhai et al. (2022); Jain et al. (2023). Closest to our work, Hernandez et al. (2021) have analyzed transfer learning but with a focus on how the cross-entropy loss behaves as the *finetuning* data scales. Unlike our work, their scaling law describes the relation between the size of a (finetuning) dataset and the cross-entropy loss on the *same* dataset – making this closer to the standard scaling laws in the literature since the finetuning loss and the finetuning dataset are computed over samples from the same distribution. On the other hand, we propose scaling laws for the *downstream* metrics on the *finetuning* dataset as the *pretraining* data scales – switching the focus to an "out-of-distribution" analysis. The only work we are aware of that has proposed scaling laws for the *downstream task performance* as a function of pretraining dataset size is by Sun et al. (2017) who have focused on classification tasks in the vision domain and used small models relative to LLMs.

Transferability metrics and value of pretraining. While it may be commonly suggested that pretraining data improves both *upstream* and *downstream* performance, this rule has been challenged in the vision domain. Zoph et al. (2020); He et al. (2019); Shen et al. (2019); Ghiasi et al. (2018); Mikami et al. (2022) have demonstrated that pretraining can sometimes have no effect on the *downstream* task performance and sometimes it can even hurt the performance. We make similar observations in the language domain with extensive experiments on LLMs and identify cases where (a) adding more pretraining data hurts the *downstream task performance* when pretraining data is not aligned enough with the task and (b) pretraining does not improve the *downstream task performance* noticeably when the finetuning dataset is large enough. Another related line of work is on transferability metrics Tamkin et al. (2020); Chiang & Lee (2022); Ibrahim et al. (2022); Tran et al. (2019); Agostinelli et al. (2022); Tran et al. (2019); Nguyen et al. (2020); You et al. (2021); Dai et al. (2019); Huang et al. (2022); Ibrahim et al. (2022); Tran et al. (2019); Bao et al. (2019); Van Asch & Daelemans (2010); Plank & Van Noord (2011), which are efficient heuristics used to select the most appropriate source models or pretraining data for a given target task. We note that transferability metrics are designed to solve *ranking* problems, different from scaling laws. For example, these metrics answer questions such as given a pool of source models (or pretraining datasets), which source model (or pretraining dataset) is the best to finetune on for a given target task. These metrics are not designed to predict the performance of the model when key quantities (e.g., pretraining data size) are scaled.

C A Guide for Pretraining Data Valuation

In this section, combining our findings on the scaling behavior of BLEU score, we propose the following guide for assessing the value of pretraining dataset for a target downstream task:

1. Given a pretraining dataset, pretrain as long as possible under the given computational and time constraints[3]. Periodically choose pretraining checkpoints, finetune on them, and record the downstream performance metric (we recommend the BLEU score over cross-entropy due to the discussion in Section 2.3).
2. Since the law in (1) has three coefficients to be fit, once we have 3 pairs of (number of pretraining tokens seen, BLEU score), we *try* to find the optimal coefficients. If the BLEU scores have a non-monotonic behavior, we cannot fit the scaling law. Since the non-monotonic behavior could be an indication of misalignment (following the discussion in Section 2.3), we recommend checking the BLEU score of the best available finetuned checkpoint and comparing it to the performance of the non-pretrained model trained on the downstream task directly. If the scaling law fits well, then we make the initial prediction for the BLEU score as we increase the pretraining dataset size (or pretrain for more steps). If we are not satisfied with the predicted BLEU score, then we conclude that it is not worth pretraining on this dataset. If the predicted BLEU score is high enough, then we keep pretraining until we reach the target BLEU score. If the scaling law breaks at any point, we conclude that the pretraining dataset is not sufficiently aligned with the downstream task and pretraining further may not be beneficial.

D Additional Details on the Experimental Setup

In the experiments, we first pretrain a model without doing more than one pass over any of the examples. Then, we finetune selected checkpoints of the pretrained model. Naturally, there is a one-to-one mapping between the checkpoint number and the number of pretraining tokens seen. This way, we collect pairs of (number of pretraining tokens, BLEU score) and (number of pretraining tokens, *downstream* cross-entropy loss) to analyze them with the proposed scaling laws in (1) and (2). All the plots are on a $\log$-$\log$ scale.

Model. We use the 3-billion encoder-decoder T5 model with 24 encoder layers, 24 decoder layers, embedding dimension 1024, and 32 heads with dimension 128. We note that this is the same model as the T5-3B model in Abnar et al. (2022). In Appendix E.1, we also provide results with a smaller 770-million encoder-decoder T5 model. This model corresponds to T5-Large in Raffel et al. (2020). We share more details about the architectures in Tables 1 and 2. For encoding the text as WordPiece tokens (Sennrich et al., 2016; Kudo, 2018), we use SentencePiece (Kudo & Richardson, 2018) trained with a vocabulary of size $250,112$ that covers all the languages in the MC4 dataset (Raffel et al., 2020).

Datasets. We use the English (en), German (de), French (fr), and Romanian (ro) portions of the MC4 dataset. We experiment with both pretraining on these languages individually as well as mixing pairs of languages. In Figure 1, we present results for the models pretrained on (*left*) a mixture of 50% en-MC4 + 50% de-MC4, (*center*) a mixture of 50% en-MC4 + 50% fr-MC4, and (*right*) a mixture of 50% en-MC4 + 50% ro-MC4 – meaning that 50% of one pretraining batch is sampled

[3]We avoid repeating sequences as repetitions may complicate the scaling behavior Hernandez et al. (2022); Muennighoff et al. (2023); Tirumala et al. (2023). This means as pretraining goes on, we effectively pretrain each checkpoint on a "larger dataset".

Table 1: T5-3B Raffel et al. (2020) architecture details.

Embedding Dimension	1024
Number of Heads	32
Number of Encoder Layers	24
Number of Decoder Layers	24
Head Dimension	128
MLP Dimension	16384

Table 2: T5-770M Raffel et al. (2020) architecture details.

Embedding Dimension	1024
Number of Heads	16
Number of Encoder Layers	24
Number of Decoder Layers	24
Head Dimension	64
MLP Dimension	2816

from en-MC4 and the other 50% is sampled from the other language. In Figure 2, we show results for the models pretrained only on en-MC4. In Figure 5, in addition to these, we also present results for the models pretrained on a mixture of 30$\%$ en-MC4 + 70$\%$-fr and a mixture of 70$\%$ en-MC4 + 30$\%$-fr as well as models pretrained only on de-MC4, only on fr-MC4, and only on ro-MC4. We finetune the pretrained models on WMT-17 en-de (Bojar et al., 2017), WMT-15 en-fr (Bojar et al., 2014), and WMT-16 en-ro (Bojar et al., 2016), separately. To understand the effect of the finetuning dataset size on the scaling laws, we sometimes use a smaller randomly sampled portion from these translation datasets and indicate the number of tokens used.

Hyperparameters. During pretraining, we use a batch size of 256 and a sequence length of 512 for $1,000,000$ steps except for the ro-MC4 pretraining. For ro-MC4, we pretrain for $510,000$ steps since otherwise, we would need to do repetitions over the sequences. Following Raffel et al. (2020), we use an "inverse square root" learning rate schedule, $\frac{1}{\sqrt{\max(n,k)}}$, where n is the current pretraining step and k is set to 10^4. We do a grid search for the base learning rate from $\{0.05, 0.1, 0.5, 1.0, 2.0, 5.0\}$ and pick the best one for each pretrained model based on *upstream* cross entropy. During finetuning, again following Raffel et al. (2020), we use a batch size of 128 and a sequence length of 512 for 300 steps. We use a constant learning rate by selecting the best from $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. In both stages, we use the AdaFactor optimizer (Shazeer & Stern, 2018).

Optimizing the scaling law coefficients. To fit the coefficients in the scaling laws in (1) and (2), similar to Hoffmann et al. (2022), we use the Huber loss (Huber, 1992) and the L-BFGS algorithm (Nocedal, 1980) to estimate the scaling law robustly in the presence of outliers. For the Huber loss, we use $\delta = 0.1$ for the BLEU score and $\delta = 1e - 3$ for the *downstream* cross-entropy loss. We select the best fit among a grid of initializations and report the prediction error computed via the Huber loss in Appendix E.2. To optimize the coefficients, we use the first four data points that require the smallest amount of pretraining data and leave the remaining data points as held-out data to evaluate the accuracy of the laws. We note that, ideally, three points should be enough since both laws have three coefficients to be optimized for. However, adding more points improves the fit by making the optimization more robust to outliers. We refer the reader to Appendix E.2 for the list of optimized coefficients and the prediction errors.

Next, we provide more details on how we optimize the coefficients of the scaling laws. Following Hoffmann et al. (2022), we use the Huber loss (Huber, 1992) to minimize overfitting to the outliers. Huber loss is particularly useful to suppress the effect of the outlier data points in the optimization problem. More specifically, if the data point with value r is predicted by the law as $\hat{r}$, the loss for that data point would be

$$\ell_\delta(r, \hat{r}) = \begin{cases} \frac{1}{2}(r - \hat{r})^2 & \text{for } |r - \hat{r}| \leq \delta, \\ \delta \cdot (|r - \hat{r}| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases} \tag{3}$$

Due to the numerical range difference between the BLEU score (between 0 and 100) and the *downstream* cross-entropy typically taking much smaller values, we use $\delta = 0.1$ for the BLEU score law in (1) and $\delta = 1e - 3$ for the *downstream* cross-entropy law in (2).

For optimization, we use the L-BFGS algorithm (Nocedal, 1980). Specifically, for the BLEU score law in (1), we solve

$$\min_{E,A,\alpha,\beta} \sum_{\text{Data point } i} \ell_\delta(\log f_i, \log \hat{f}(D_{pi})), \tag{4}$$

where D_{pi} is the pretraining dataset size and f_i is the BLEU score for the data point i, and $\hat{f}(\cdot)$ is the approximation for the optimal law $f(\cdot)$. Similarly, for the *downstream* cross-entropy loss law in (2), we solve

$$\min_{E,A,\alpha} \sum_{\text{Data point } i} \ell_\delta(\log L_i, \log \hat{L}(D_{pi})), \tag{5}$$

where D_{pi} is the pretraining dataset size and L_i is the *downstream* cross-entropy loss for the data point i, and $\hat{L}(\cdot)$ is the approximation for the optimal law $L(\cdot)$.

E Results and Analysis

In Figure 1, we analyze the models that are pretrained on different portions of *(left)* a mixture of 50% en-MC4 + 50% de-MC4, *(center)* a mixture of 50% en-MC4 + 50% fr-MC4, and *(right)* a mixture of 50% en-MC4 + 50% ro-MC4. These models are then finetuned on different portions of *(left)* en-de, *(center)* en-fr, and *(right)* en-ro translation datasets. In the top row, we report the BLEU score and, in the bottom row, we report the *downstream* cross-entropy loss. The dotted, dashed, and solid lines correspond to the scaling laws in (1) and (2) for different finetuning dataset sizes D_f. The black lines correspond to “non-pretrained” models (randomly initialized) that are directly trained on different portions of the finetuning dataset. In all cases, the scaling laws fit well to the empirical results (the markers) with prediction error at most 0.061 for the BLEU score ($\delta = 0.1$) and $5.95e - 12$ for the *downstream* cross-entropy ($\delta = 1e - 3$) (see Appendix E.2 for more details). As expected, as the finetuning dataset size increases (e.g., going in the order of dotted-dashed-solid lines), the BLEU score increases and the cross-entropy loss decreases smoothly and monotonically. Similarly, as the pretraining dataset size D_p increases (along the x-axis), we see improvements in both metrics. Notice that the improvements by an increase in the pretraining dataset size is more effective for smaller finetuning datasets. When the finetuning dataset is large enough (e.g., solid lines), BLEU score is more or less constant regardless of the pretraining dataset size. In fact, we see little to no improvement of pretraining compared to the non-pretrained models (black lines) when the finetuning dataset is large. **This implies that, for these tasks, there is no need to pretrain the models when the finetuning dataset is large enough. Luckily, we can correctly predict whether this is going to be the case (i.e., whether the available finetuning data is enough to eliminate pretraining altogether) with the use of scaling laws. All we need to do is to pretrain the model on a small portion of the pretraining dataset with reasonable compute cost to optimize the coefficients of the scaling laws, and then follow the guideline provided in Section C.**

In Figure 2, we change the pretraining dataset to 100% en-MC4 in all plots. Intuitively, we expect this dataset to be less aligned with the translation tasks than the multilingual pairs in Figure 1 since it does not include one of the languages in the translation tasks. Indeed, we see smaller BLEU score and higher cross-entropy loss in general for the same finetuning dataset size. Most of the conclusions from Figure 1 carry over to the results in Figure 2. For instance, the pretraining data matters less when the finetuning dataset is large enough. One noticeable difference is in the BLEU scores for the en-fr

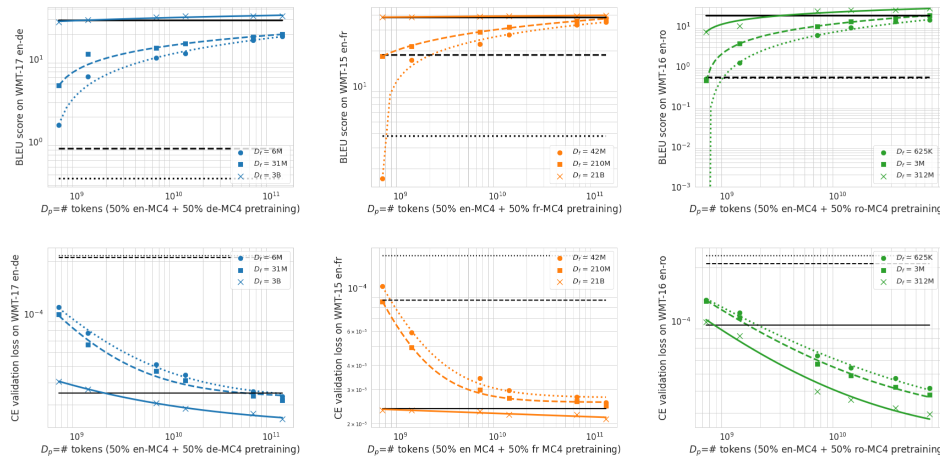

Figure 3: **(top) BLEU score vs pretraining dataset size: $\mathbf{f(D_p)} = (\log(\mathbf{A} \cdot \mathbf{D_p^{\alpha}}))^{\beta}$.** ***(left)*** WMT-17 en-to-de translation task. Pretraining dataset has 50% en-MC4 + 50% de-MC4. Dotted, dashed, and solid blue curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 6M$, $D_f = 31M$, $D_f = 3B$ tokens, respectively. ***(center)*** WMT-15 en-to-fr translation task. Pretraining dataset has 50% en-MC4 and 50% fr-MC4. Dotted, dashed, and solid orange curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 42M$, $D_f = 210M$, $D_f = 21B$ tokens, respectively. ***(right)*** WMT-16 en-to-ro translation task. Pretraining dataset has 50% en-MC4 + 50% ro-MC4. Dotted, dashed, and solid green curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 625K$, $D_f = 3M$, $D_f = 312M$ tokens, respectively. **(bottom) Cross-entropy (CE) validation loss vs pretraining dataset size: $\mathbf{L(D_p)} = \mathbf{E} + \frac{\mathbf{A}}{\mathbf{D_p^{\alpha}}}$.** Same models as the top row. For all the plots, the markers are the actual experimental results and the black horizontal curves correspond to the non-pretrained model directly trained on the task dataset. **The finetuning dataset size increases in the order of dotted-dashed-solid for all the curves including the black horizontal lines.**

translation task (*center*). We see that, for $D_f = 42M$ and $D_f = 210M$, the scaling law for BLEU score actually breaks once the pretraining dataset size passes a threshold while the cross-entropy loss scales as expected. This is counter-intuitive because the BLEU score sometimes decreases for larger pretraining dataset. Notice that this break in scaling law does not happen in en-de or en-ro translation tasks as the scaling law fits well to the pretraining data with prediction error at most 0.025 for these tasks ($\delta = 0.1$). To better investigate this, in Figure 5, we take a closer look at some less aligned pretraining datasets due to the choice of language.

In Figure 5-(*left*), we provide the scaling laws for en-de translation task where the pretraining datasets are 100% en-MC4 (same as Figure 2-(*left*)), 50% en-MC4 and 50% de-MC4 (same as Figure 1-(*left*)), 100% de-MC4, 100% fr-MC4 (less aligned), and 100% ro-MC4 (less aligned). Notice that the last two pretraining datasets are expected to be the least aligned with the translation task since the translation pair does not include these languages. We see that, despite this, the scaling laws consistently fit well for both the BLEU score and the cross-entropy loss. However, this is not always the case for the en-fr translation task. In Figure 5-(*right*), we provide the scaling laws for the en-fr translation task where the pretraining datasets are different mixtures of en-MC4 and fr-MC4 datasets. We also include the "less aligned" pretraining datasets such as 100% de-MC4 and 100% ro-MC4. Surprisingly, we see that the scaling law for the BLEU score breaks after some point for the only-English (100% en-MC4), only-German (100% de-MC4), and only-Romanian (100% ro-MC4) pretraining datasets while the cross-entropy loss always follows the scaling law in (2). Interestingly, we do not observe such a break in the BLEU score scaling for the only-French (100% fr-MC4) pretraining dataset – hinting that not including French data in pretraining leads to poor scaling in the en-fr translation task but not including English does not have such an effect. We also notice that the BLEU score is the lowest for these three pretraining datasets where scaling breaks. **This suggests that the scaling law in (1) works well for the BLEU score as long as the pretraining dataset has the promise to give rise to a good performance. However, when the scaling law does not fit well, we may suspect the**

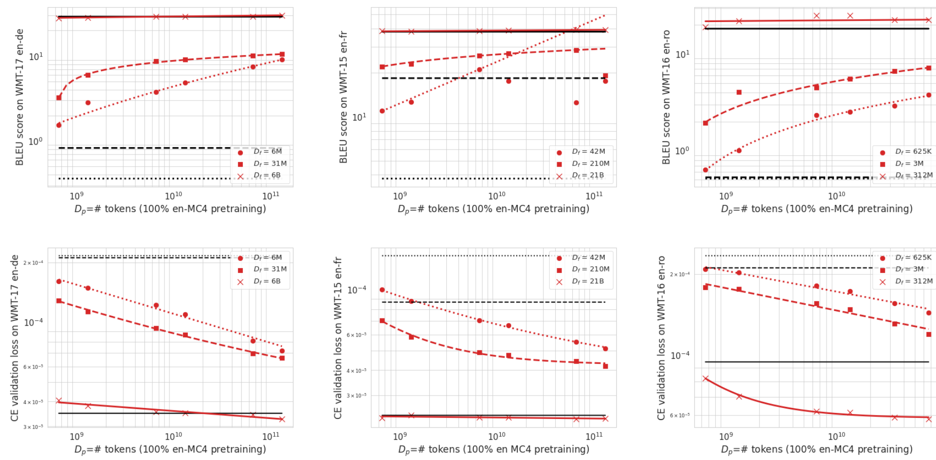

Figure 4: **(top) BLEU score vs pretraining dataset size: $\mathbf{f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}}$.** ***(left)*** WMT-17 en-to-de translation task. Dotted, dashed, and solid red curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 6M$, $D_f = 31M$, $D_f = 3B$ tokens, respectively. ***(center)*** WMT-15 en-to-fr translation task. Dotted, dashed, and solid red curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 42M$, $D_f = 210M$, $D_f = 21B$ tokens, respectively. ***(right)*** WMT-16 en-to-ro translation task. Dotted, dashed, and solid red curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 625K$, $D_f = 3M$, $D_f = 312M$ tokens, respectively. **(bottom) Cross-entropy (CE) validation loss vs pretraining dataset size: $\mathbf{L(D_p) = E + \frac{A}{D_p^{\alpha}}}$.** Same models as the top row. For all the plots, the markers are the actual experimental results and the black horizontal curves correspond to the non-pretrained model directly trained on the task dataset. **The finetuning dataset size increases in the order of dotted-dashed-solid for all the curves including the black horizontal lines.**

BLEU score to be low overall. Therefore, whether we can fit the scaling law for the BLEU score seems to give a good indication about the degree of alignment between the pretraining data and the particular translation task.

Remark 1. We observe another interesting phenomenon in Figure 5. For both en-de and en-fr tasks, 100% en-MC4 leads to significantly worse BLEU score and *downstream* cross-entropy than the more aligned 50% en-MC4 + 50% de/fr-MC4 balanced datasets, respectively. However, de-MC4 and fr-MC4 perform almost as well as the balanced datasets in en-de and en-fr tasks. We leave the investigation of why pretraining on only German/French helps more than pretraining on only English for the given en-de and en-fr tasks to future work.

We also highlight that we cannot make any strong conclusion about the degree of alignment of the pretraining dataset with the task by only looking at the *downstream* cross-entropy loss because of the inconsistency with the BLEU score, a task-related metric, observed in the en-fr plots in Figures 2 and 5. This is a counter-example for the claim by Gordon et al. (2021) that the two metrics have an exponential relation. To better demonstrate this, in Figure 6, we provide a BLEU score vs. *downstream* cross-entropy log-log plot for en-de and en-fr translation tasks, respectively. While the two metrics indeed seem correlated in Figure 6-(***left***) on the en-de task, we observe a somewhat arbitrary relation for the en-fr task in Figure 6-(***right***) in some cases – which clearly cannot be explained with an exponential relation. **This suggest that *downstream* cross-entropy is not always a good indicator for BLEU score. This raises the question whether the scaling laws that have been developed for the *upstream* cross-entropy loss are actually useful predictors for models' downstream behavior.**

Remark 2. We also revisit the definition of the BLEU score to better understand the root cause of the non-smooth behavior and check if we could see a smooth monotonic scale in at least some elements

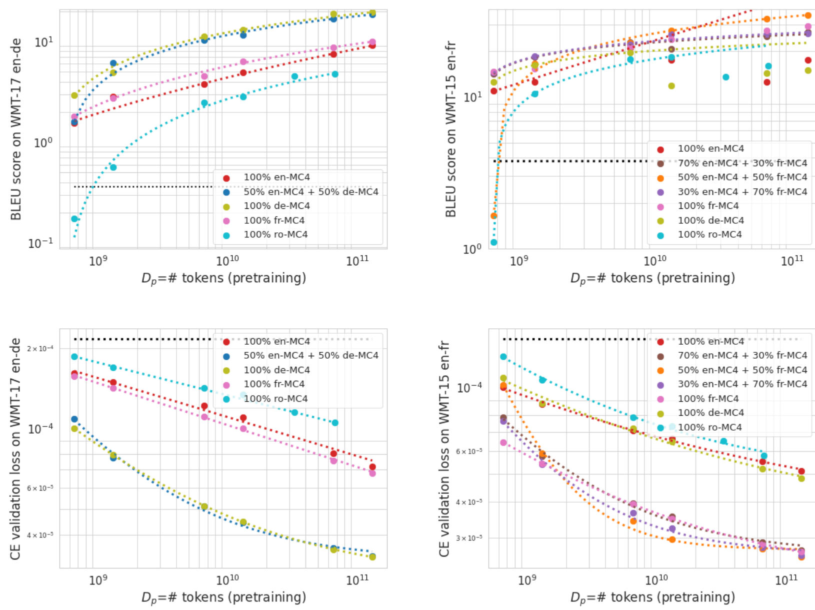

Figure 5: Comparison of scaling behavior for different pretraining datasets. **(top) BLEU score vs pretraining dataset size: $\mathbf{f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}}$.** ***(left)*** WMT-17 en-de translation task. ***(right)*** WMT-15 en-fr translation task. **(bottom) Cross-entropy (CE) validation loss vs pretraining dataset size: $\mathbf{L(D_p) = E + \frac{A}{D_p^{\alpha}}}$.** Same as the top row but for CE loss instead of BLEU score. For all the plots, the markers are the actual experimental results and the black horizontal curves correspond to the non-pretrained model directly trained on the task dataset.

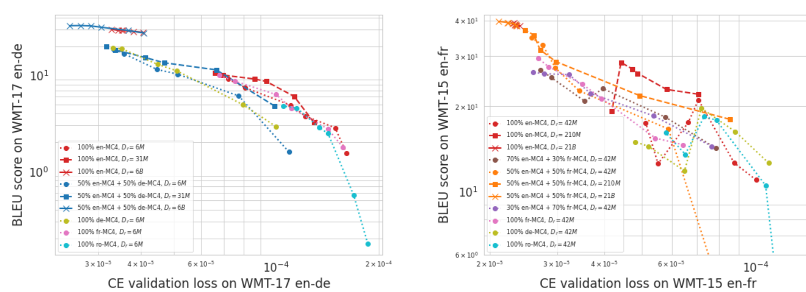

Figure 6: **BLEU score vs. *downstream* cross-entropy loss.** ***(left)*** For en-de translation task, we see a consistent correlation between the two metrics for all the pretraining datasets. This supports the findings of Gordon et al. (2021). ***(right)*** For en-fr translation task, the two metrics usually show an arbitrary relation. Sometimes, the BLEU score increases while the cross-entropy also increases. Unlike the en-de results in (left), the exponential relation in (Gordon et al., 2021) is not observed here.

of the BLEU score calculation. Recall that the common form of BLEU score is defined as

$$\text{BLEU} = \text{brevity-penalty} \cdot \left(\prod_{i=1}^{4} \text{precision}_i \right)^{1/4}, \tag{6}$$

where precision_n refers to the precision of n-grams, and the second term is the geometric mean of the precision when n is varied from 1 to 4. In all the experiments, we observe brevity-penalty $= 1$, i.e., the non-smooth behavior can be attributed to the precision terms. Hence, our findings, including the scaling law in (1), would also apply for precision–another *downstream* task metric.

E.1 Additional Results (on T5-770M)

In Figures 7 and 8, we present results similar to Figures 1 and 2 in Section 4, but for T5-770M instead of T5-3B. In general, we observe a similar trend. The proposed scaling laws describe the downstream behavior well when the pretraining and downstream data are aligned.

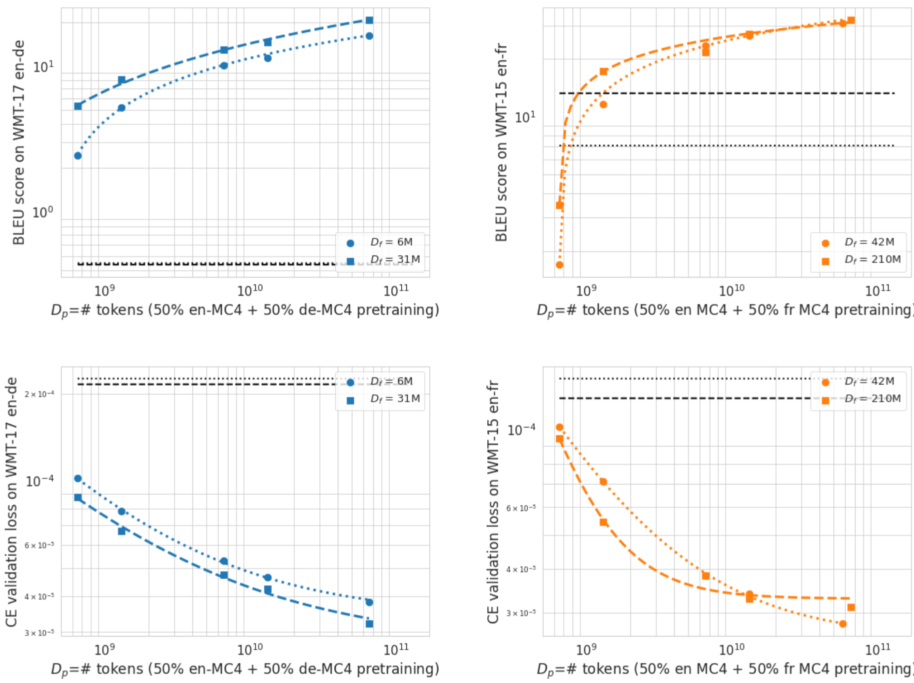

Figure 7: **(top) BLEU score vs pretraining dataset size:** $\mathbf{f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}}$. ***(left)*** WMT-17 en-to-de translation task. Pretraining dataset has 50% en-MC4 + 50% de-MC4. Dotted and dashed blue curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 6M$ and $D_f = 31M$ tokens, respectively. ***(right)*** WMT-15 en-to-fr translation task. Pretraining dataset has 50% en-MC4 and 50% fr-MC4. Dotted and dashed orange curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 42M$ and $D_f = 210M$ tokens, respectively. **(bottom) Cross-entropy (CE) validation loss vs pretraining dataset size:** $\mathbf{L(D_p) = E + \frac{A}{D_p^{\alpha}}}$. Same models as the top row. For all the plots, the markers are the actual experimental results and the black horizontal curves correspond to the non-pretrained model directly trained on the task dataset. **The finetuning dataset size increases in the order of dotted-dashed for all the curves including the black horizontal lines.**

E.2 Optimized Coefficients and Prediction Errors of the Scaling Laws

In Tables 3, 4, 5, and 6, we provide the optimized coefficients for the scaling laws plotted in Figures 1 and 2 together with the prediction error.

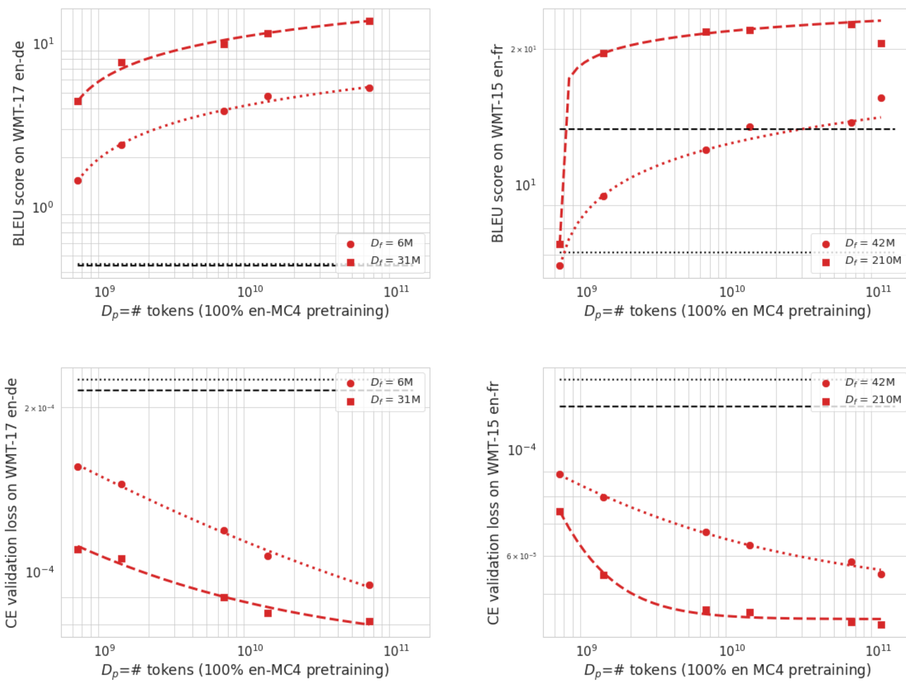

Figure 8: **(top) BLEU score vs pretraining dataset size: $\mathbf{f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}}$.** ***(left)*** WMT-17 en-to-de translation task. Dotted and dashed red curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 6M$ and $D_f = 31M$ tokens, respectively. ***(right)*** WMT-15 en-to-fr translation task. Dotted and dashed red curves correspond to the fitted scaling laws for different finetuning dataset sizes, $D_f = 42M$ and $D_f = 210M$ tokens, respectively. **(bottom) Cross-entropy (CE) validation loss vs pretraining dataset size: $\mathbf{L(D_p) = E + \frac{A}{D_p^{\alpha}}}$.** Same models as the top row. For all the plots, the markers are the actual experimental results and the black horizontal curves correspond to the non-pretrained model directly trained on the task dataset. **The finetuning dataset size increases in the order of dotted-dashed for all the curves including the black horizontal lines.**

Table 3: The coefficients for the BLEU score law $f(D_p) = (\log(A \cdot D_p^{\alpha}))^{\beta}$ for the results in Figure 1-(**top**). For the BLEU score laws, we use $\delta = 0.1$ for the Huber Loss. We report $\log A$ instead of A since A typically takes very small and very large values.

Pretraining Dataset	Finetuning Dataset	Finetuning Dataset Size	$\log A$	α	β	Prediction Error
50% en + 50% de-MC4	WMT-17 en-de	6M	-180.75	9.00	0.75	0.034
50% en + 50% de-MC4	WMT-17 en-de	31M	-1.68×10^3	84.04	0.49	0.050
50% en + 50% de-MC4	WMT-17 en-de	3B	-1.64×10^8	9.91×10^6	0.19	0.048
50% en + 50% fr-MC4	WMT-15 en-fr	42M	-1.82×10^4	8.98×10^2	0.42	0.061
50% en + 50% fr-MC4	WMT-15 en-fr	210M	-2.33×10^4	1.21×10^3	0.40	0.013
50% en + 50% fr-MC4	WMT-15 en-fr	21B	5.08×10^3	4.61×10^8	0.16	0.005
50% en + 50% ro-MC4	WMT-16 en-ro	625K	-36.02	1.77	1.28	0.042
50% en + 50% ro-MC4	WMT-16 en-ro	3M	$-0.115.03$	5.69	0.89	0.015
50% en + 50% ro-MC4	WMT-16 en-ro	312M	-1.82×10^4	9.04×10^2	0.40	0.015

Table 4: The coefficients for the *downstream* cross-entropy law $L(D_p) = E + \frac{A}{D_p^\alpha}$ for the results in Figure 1-(**bottom**). For the *downstream* cross-entropy laws, we use $\delta = 10^{-5}$ for the Huber Loss.

Pretraining Dataset	Finetuning Dataset	Finetuning Dataset Size	E	A	α	Prediction Error
50% en + 50% de-MC4	WMT-17 en-de	6M	3.21×10^{-5}	35.45	0.64	1.36×10^{-12}
50% en + 50% de-MC4	WMT-17 en-de	31M	3.28×10^{-5}	4.70×10^{2}	0.78	3.17×10^{-12}
50% en + 50% de-MC4	WMT-17 en-de	3B	2.24×10^{-5}	2.56×10^{-2}	0.36	5.76×10^{-14}
50% en + 50% fr-MC4	WMT-15 en-fr	42M	2.72×10^{-5}	2.01×10^{6}	1.18	7.52×10^{-13}
50% en + 50% fr-MC4	WMT-15 en-fr	210M	2.57×10^{-5}	1.75×10^{7}	1.30	2.24×10^{-13}
50% en + 50% fr-MC4	WMT-15 en-fr	21B	1.11×10^{-7}	3.41×10^{-5}	1.82×10^{-2}	5.20×10^{-14}
50% en + 50% ro-MC4	WMT-16 en-ro	625K	2.45×10^{-5}	0.49	0.41	3.61×10^{-12}
50% en + 50% ro-MC4	WMT-16 en-ro	3M	2.62×10^{-5}	2.40	0.49	2.19×10^{-12}
50% en + 50% ro-MC4	WMT-16 en-ro	312M	2.08×10^{-5}	3.94	0.53	5.95×10^{-12}

Table 5: The coefficients for the BLEU score law $f(D_p) = (\log(A \cdot D_p^\alpha))^\beta$ for the results in Figure 2-(**top**). For the BLEU score laws, we use $\delta = 0.1$ for the Huber Loss. We report $\log A$ instead of A since A typically takes very small and very large values.

Pretraining Dataset	Finetuning Dataset	Finetuning Dataset Size	$\log A$	α	β	Prediction Error
100% en-MC4	WMT-17 en-de	6M	-1.88	0.15	3.30	0.014
100% en-MC4	WMT-17 en-de	31M	-1.81×10^{4}	896.12	0.28	0.006
100% en-MC4	WMT-17 en-de	3B	1.02×10^{-7}	104.92	0.42	0.015
100% en-MC4	WMT-15 en-fr	42M	1.00	2.57×10^{-5}	1.11×10^{4}	0.042
100% en-MC4	WMT-15 en-fr	210M	-6.38×10^{7}	3.43×10^{6}	0.20	0.034
100% en-MC4	WMT-15 en-fr	21B	204.81	3.80×10^{14}	9.97×10^{-3}	0.004
100% en-MC4	WMT-16 en-ro	625K	-10.54	0.55	1.12	0.008
100% en-MC4	WMT-16 en-ro	3M	-40.41	2.11	0.79	0.025
100% en-MC4	WMT-16 en-ro	312M	3.61	8.17×10^{5}	0.19	0.018

Table 6: The coefficients for the *downstream* cross-entropy law $L(D_p) = E + \frac{A}{D_p^\alpha}$ for the results in Figure 2-(**bottom**). For the *downstream* cross-entropy laws, we use $\delta = 10^{-5}$ for the Huber Loss.

Pretraining Dataset	Finetuning Dataset	Finetuning Dataset Size	E	A	α	Prediction Error
100% en-MC4	WMT-17 en-de	6M	3.22×10^{-13}	3.18×10^{-3}	0.15	5.79×10^{-12}
100% en-MC4	WMT-17 en-de	31M	3.24×10^{-5}	5.20×10^{-3}	0.20	9.25×10^{-13}
100% en-MC4	WMT-17 en-de	3B	2.24×10^{-5}	2.56×10^{-2}	0.36	5.76×10^{-14}
100% en-MC4	WMT-15 en-fr	42M	3.49×10^{-5}	1.05×10^{-2}	0.25	3.63×10^{-13}
100% en-MC4	WMT-15 en-fr	210M	4.24×10^{-5}	19.39	0.66	5.40×10^{-13}
100% en-MC4	WMT-15 en-fr	21B	1.26×10^{-7}	2.59×10^{-5}	4.81×10^{-3}	3.63×10^{-14}
100% en-MC4	WMT-16 en-ro	625K	5.79×10^{-12}	1.03×10^{-3}	7.76×10^{-2}	5.56×10^{-12}
100% en-MC4	WMT-16 en-ro	3M	1.78×10^{-12}	9.98×10^{-4}	8.33×10^{-2}	8.23×10^{-12}
100% en-MC4	WMT-16 en-ro	312M	5.85×10^{-5}	1.37×10^{3}	0.88	3.05×10^{-13}