Language-Independent Representations Improve Zero-Shot Summarization

Vladimir Solovyev* **Danni Liu** **Jan Niehues**
Karlsruhe Institute of Technology, Germany
vladimir.solovyev.90@gmail.com, {danni.liu, jan.niehues}@kit.edu

Abstract

Finetuning pretrained models on downstream generation tasks often leads to catastrophic forgetting in zero-shot conditions. In this work, we focus on summarization and tackle the problem through the lens of language-independent representations. After training on monolingual summarization, we perform zero-shot transfer to new languages or language pairs. We first show naively finetuned models are highly language-specific in both output behavior and internal representations, resulting in poor zero-shot performance. Next, we propose query-key (QK) finetuning to decouple task-specific knowledge from the pretrained language generation abilities. Then, after showing downsides of the standard adversarial language classifier, we propose a balanced variant that more directly enforces language-agnostic representations. Moreover, our qualitative analyses show removing source language identity correlates to zero-shot summarization performance. Our code is openly available[1].

1 Introduction

Pretrained multilingual models (Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021; Lin et al., 2022) have been established as promising sources of transfer learning, where task-specific finetuning benefits from the general knowledge learned on diverse unsupervised data. However, due to data or computational constraints, the task-specific data often only cover a limited subset of the languages in pretraining. Therefore, during finetuning it is crucial to retain the knowledge of the pretrained model and to enable zero-shot transfer, i.e., performing the task on more languages covered by the pretrained model. While zero-shot crosslingual transfer has shown very promising results on sequence classification or labeling problems (Pires et al., 2019; Conneau and Lample, 2019; Wu and Dredze, 2019), it remains challenging for generation tasks (Rönnqvist et al., 2019; Vu et al., 2022; Li and Murray, 2023) including summarization and translation. A main obstacle is catastrophic forgetting (French and Chater, 2002), where languages supported by the pretrained model but not covered in the finetuning data are forgotten. In this work, we use summarization as a testbed for various types of zero-shot generation. As shown in Figure 1, given a pretrained model and intralingual summarization training data in some languages (A→A, B→B), we aim for zero-shot *intralingual* and *crosslingual* summarization on new languages (C→C) and language pairs (A→B, A→C) respectively.

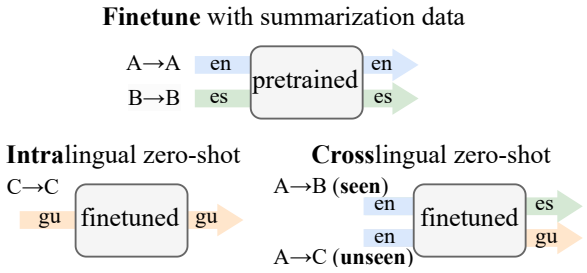

Figure 1: We finetune a pretrained model (e.g. mBART) on intralingual summarization data and test it on zero-shot intralingual and crosslingual summarization.

To alleviate catastrophic forgetting, one line of work trains on additional unsupervised data (Maurya et al., 2021; Vu et al., 2022; Chronopoulou et al., 2023). Besides the computational overhead, this approach raises a theoretical question: As the pretrained language model has already learned extensively on unsupervised data, is it necessary to re-learn language modeling in task-specific finetuning? We therefore explore a more challenging case of only using paired summarization data without relying on any unsupervised data.

We identify two challenges when generalizing summarization abilities to new languages. First, decoupling the *task*-specific knowledge from the

*Work done while at Karlsruhe Institute of Technology
[1]https://github.com/vladsolovyev/fairseq_summarization/tree/main/summarization_scripts

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 772–782
June 16-21, 2024

train on **en**			train on en+es+ru				
es-es	**ru-ru**	**gu-gu**	**gu-gu**	**es-en**	**ru-en**	**es-ru**	**tr-en**
0.2	2.3	13.4	99.6	0.0	0.0	0.0	1.3

Table 1: Proportion of generated summaries in the correct language (%) under zero-shot conditions. Codes: es (Spanish), ru (Russian), gu (Gujarati), tr (Turkish).

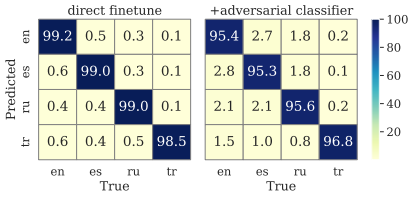

Figure 2: Accuracy of a probing classifier. Higher accuracy indicates more language-specific representations.

language generation abilities is essential. In response, we propose a new finetuning method based on query and keys, which is shown effective for both intralingual and crosslingual zero-shot setting. For crosslingual zero-shot settings, it is also crucial to decouple *language* from *content*, i.e., creating language-agnostic representations. This has been shown to facilitate zero-shot crosslingual generation in general (Pham et al., 2019; Wu et al., 2022; Duquenne et al., 2023). There a prominent approach is adversarial training (Ganin et al., 2016; Chen et al., 2018), where the model is trained to deceive a language classifier. We show the existing formulation fails to fully achieve language-agnostic representations, and improves it by explicitly incentivizing the model to deceive the classifier into a uniform class distribution.

2 Current Models are Highly Language-Specific

We first show that naive finetuning makes the models highly language-specific in their *output behavior* and *internal representations*. Table 1 shows the proportion of outputs in the correct language after finetuning mBART on intralingual summarization data. Finetuning on English only leads the model to forget its generation ability for other languages, resulting in *off-target generation* (Zhang et al., 2020a; Pfeiffer et al., 2023), where a wrong target language, often one with supervised data, is generated. Although multilingual training largely resolves off-target generation in intralingual settings[2], the problem persists for crosslingual generation. As zero-shot crosslingual generation relies on language-agnostic representations, we test for this with a probing analysis (Adi et al., 2017). Specifically, we assess the difficulty of recovering the source language identity on the encoder output. Given a trained model, we train a token-level classifier for the input languages on the encoder outputs.[3] As shown in Figure 2, the classifier can almost perfectly recover the source language. Even after explicitly encouraging language-agnostic representations with an adversarial language classifier (Arivazhagan et al., 2019), recovering the source language identity remains easy.

3 Approaches

3.1 Decoupling Language from Task

Query-Key (QK) Finetuning Prior works on zero-shot generation (Chi et al., 2020; Maurya et al., 2021; Li et al., 2021) have highlighted the need for selective finetuning to mitigate forgetting, where the consensus is updating the encoder and cross-attention weights only. However, existing methods treat attention weights as a whole. A closer look at the attention module reveals that, only the value projections determine the basis of the upcoming transformations, whereas the query and key control how the inputs are aggregated. We hypothesize that the value projections should be kept unchanged to prevent losing pretrained generation capabilities during finetuning. In contrast, query and key are updated as adaptation to specific tasks. Therefore, we propose a selective finetuning approach, which only updates the query and key projection weights of encoder self-attention and cross-attention.[4]

Two-Step Finetuning For the more challenging case of crosslingual zero-shot summarization, our approach is motivated by the fact that the task consists of two subtasks: translation and summarization. We first finetune the pretrained model for translation[5]. Then we finetune again on intralingual summarization using our proposed query-key

[2]This is consistent with recent or concurrent findings (Chirkova et al., 2023; Pfeiffer et al., 2023).

[3]Details on the probing analysis are in Appendix A.

[4]It is extendable to the parameter-efficient finetuning (PEFT) approach LoRA (Hu et al., 2022) by placing the adapters on the query and key weights only. In the experiments we do not compare to prominent PEFT approaches like prompt tuning (Lester et al., 2021) and LoRA, as prior works have shown they in their standard forms still suffer from catastrophic forgetting in finetuning (Vu et al., 2022) or continual pretraining (Li and Lee, 2024).

[5]We do not use the finetuned mBART on translation (Tang et al., 2020) as it can only translate from or into English.

finetuning to retain its crosslingual capabilities acquired from translation in the first step.

3.2 Decoupling Language from Content

An adversarial language classifier is often used to decouple language from the semantic representations of input contents. Most existing works use the cross-entropy loss (Arivazhagan et al., 2019; Mallinson et al., 2020) and a gradient reversal layer (Ganin et al., 2016) to update the encoder weights in the opposite direction of the classifier accuracy.[6] A problem with the cross-entropy-based formulation is that it operates on *single classes* and does not incentivize language-agnostic representations on the output *distribution* level. The adversarial classifier could potentially be shift all its predicted probability mass to another language, achieving a low classification accuracy but leaving the representations still language-specific. Indeed as shown in Figure 2, even after training with this objective, a probing classifier can still easily learn to recover the source language identity.

Balanced Adversarial Language Classifier Given the drawback above, we propose a balanced adversarial objective. Specifically, we train the encoder such that a language classifier is only able to predict an uniform distribution. We achieve this by a modified adversarial loss based on the KL-divergence between the classifier output distribution and a uniform distribution:

$$\mathcal{L}_{\text{balanced\_adversarial}} = D_{\text{KL}}(P_{\theta_{\text{classifier}}} \parallel U), \quad (1)$$

where P is the classifier output distribution on token level and $U = (\frac{1}{N}, \ldots, \frac{1}{N})$ with N being the number of languages to classify.

Residual Drop We further combine our approach with residual drop (Liu et al., 2021), a method proposed for machine translation that drops the residual connection of a middle encoder layer to reduce source language signals in the encoder output.

4 Experiments and Results

4.1 Experimental Setup

Datasets We train on intralingual summarization data in English or {English, Spanish, Russian}. We use XL-Sum (Hasan et al., 2021) and WikiLingua (Ladhak et al., 2020) for experiments in Table 2 and Table 3 and respectively. The dataset details are in Appendix C.1. For the two-step finetuning, the translation data details are in Appendix C.2.

Data Conditions Besides the direct zero-shot condition, we compare to the following two data conditions:

- **Pipeline** approach translating into and from English: learn summarization on English only and translate with NLLB-200 (NLLB Team et al., 2022), a recent open multilingual translation model. Here we rely on English-only summarization as English has the most training data in both datasets, which presumably yields the highest summarization quality. While this approach ensures that the outputs are in the right language, the downsides are inference latency and translation error propagation.
- **Supervised**: train on supervised data for the zero-shot directions as performance upper-bounds.

Baselines We compare our QK finetuning to:

- Encoder finetuning (Chi et al., 2020): It only updates the encoder weights to retain the pretrained generation capability, as the decoder is expected to be more responsible for generation.
- Layernorm and attention (LNA) finetuning (Li et al., 2021): It only finetunes: 1) layernorm, 2) encoder self-attention, and 3) cross-attention.

We also compare to the standard formulation of the adversarial language classifier (Arivazhagan et al., 2019) based on the cross-entropy loss.

Training and Evaluation We initialize from the mBART (Liu et al., 2020) model, which was pretrained on monolingual data of 25 languages. Further training details are in Appendix D. To assess summarization quality, we use ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020b). We report ROUGE-L in the main text supply ROUGE-1/2 in Appendix E. We use BERTScore F_1 (F_{BERT}) following the authors' suggestions (Zhang et al., 2020b). To measure the percentage of outputs in the correct languages, we use a language identifier (Lui and Baldwin, 2012) and report accuracy (%).

4.2 Impact of Query-Key Finetuning

The intralingual zero-shot results are in Table 2 with detailed scores in Appendix E. Full finetuning (row (1)) on English-only data causes severe forgetting, where most of the output are in the wrong language, which further harms summarization scores.

[6] More details in Appendix B

ID	Model	es (train on en)		ru (train on en)		gu (train on en)		gu (train on en+es+ru)	
		RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}
(1)	Full ft.	5.4	66.0	1.0	64.3	1.2	59.1	15.1	71.8
(2)	Encoder ft. (Chi et al., 2020)	18.4	70.8	22.7	73.2	14.5	71.7	15.3	72.2
(3)	"LNA" ft. (Li et al., 2021)	20.9	71.9	21.6	72.7	10.5	68.6	16.0	72.6
(4)	Query-key ft. (ours)	**21.3**	**72.3**	**23.4**	**73.6**	**16.6**	**73.2**	**16.5**	**73.1**
(5)	Pipeline (translate to/from en)	20.7	72.1	20.2	72.4	13.6	72.1	13.6	72.1
(6)	Supervised	25.0	74.0	27.5	75.1	19.3	74.2	19.3	74.2

Table 2: Zero-shot intralingual summarization results on XL-Sum.

ID	Model	es-en		ru-en		es-ru		avg. seen		tr-en		en-tr		tr-tr		avg. unseen	
		RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}	RG-L	F_{BERT}
(1)	Baseline zero-shot	2.2	67.8	0.7	63.3	0.6	64.6	1.2	65.2	4.6	62.9	2.5	60.9	18.0	71.5	8.4	65.1
(2)	Adv. classifier	26.7	76.1	25.3	75.7	14.1	72.5	22.0	74.8	26.1	75.2	2.5	60.9	5.2	62.8	11.3	66.3
(3)	Balanced adv. (ours)	27.2	76.4	25.6	75.8	14.3	72.8	22.4	75.0	26.6	75.5	2.6	60.9	3.2	61.1	10.8	65.8
(4)	(3)+ residual drop	27.6	**76.6**	**26.3**	**76.1**	**14.8**	73.1	**22.9**	**75.3**	25.7	75.2	2.5	61.0	2.3	60.8	10.2	65.7
(5)	Two-step + QK ft. (ours)	**27.7**	76.5	**26.3**	**76.1**	**14.8**	**73.4**	**22.9**	**75.3**	**30.7**	**77.4**	**16.7**	**71.3**	**18.4**	**72.0**	**21.9**	**73.6**
(6)	Pipeline	31.1	78.1	28.5	77.3	14.4	73.8	24.7	76.4	34.1	78.7	18.7	73.1	18.5	73.2	26.3	75.0
(7)	Supervised	31.4	78.1	29.4	77.5	18.0	75.2	26.3	76.9	34.5	78.8	20.7	73.2	26.2	75.4	27.1	75.8

Table 3: Zero-shot crosslingual summarization results on WikiLingua after training on {en, es, ru} intralingual data, grouped by *seen* (new combinations of languages seen in finetuning) and *unseen* (languages not in finetuning).

QK finetuning outperforms previous methods and pipeline approach: The proposed QK finetuning in row (4) surpasses the two previous methods in row (2) and (3) by 0.4-2.1 ROUGE. It is also the only approach consistently outperforming the translation-based pipeline in row (5). Moreover, the gap to the pipeline approach magnifies from high- to low-resource languages: For es, ru, gu, our QK finetuning leads by 2.9%, 15.8%, and 22.1% ROUGE respectively. This suggests that the two translation steps in the pipeline accumulates error that harms summarization quality, especially on lower-resource languages. Compared to the oracle condition with full supervised data (row (6)), the strongest zero-shot scores with our approach lies 2.7-4.1 ROUGE behind. Given the difficulty of creating summarization data, this relatively small gap shows the potential of the zero-shot approach.

Comparison to multilingual training: Comparing the zero-shot results on Gujarati (gu), training multilingually on en+es+ru instead of English alone clearly prevents forgetting. Even full finetuning in row (1) almost always generates the correct target language. Yet, QK finetuning still surpasses rows (1)-(3). Moreover, its scores on gu when training on English only match those with multilingual training, suggesting it is a more data-efficient approach that does not rely on multilingual data.

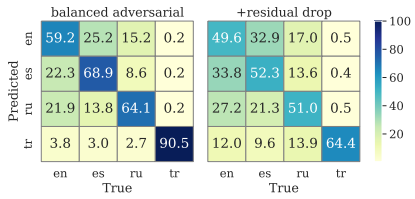

Figure 3: The models from rows (3) and (4) of Table 3 have more language-agnostic representations and stronger zero-shot performance than those in Figure 2.

4.3 Impact of Removing Language Signals

Despite its effectiveness so far, QK finetuning alone is not sufficient in crosslingual zero-shot settings. The summarization scores[7] are very low in general as a result of off-target generation. This leads to our next improvements on language-agnostic representations with results in Table 3 with detailed scores in Appendix E.

Removing language signals improves zero-shot performance for languages seen in finetuning: On language pairs where both the source and target are seen in finetuning (es-en, ru-en, es-ru), our balanced adversarial classifier in row (3) surpasses row (2) by 0.4 ROUGE on average. Combining it with the residual drop brings a further gain of 0.5 ROUGE. However, these approaches are less ef-

[7] details in Appendix E

fective on languages absent in the finetuning stage, particularly on new target languages, as shown by the poor scores on en-tr and tr-tr. Particularly, the intralingual summarization (tr-tr) quality degrades below the baseline. This shows that, to work on unseen languages, language-agnostic representations must be strengthened in target language generation.

Balanced adversarial classifier creates more language-agnostic representations: We probe the source language identity on the models trained with our balanced adversarial classifier and its combination with residual drop in Figure 3. Compare to Figure 2, these two models' representations are clearly are language-agnostic. The results show that language-agnostic representations are correlated to zero-shot cross-lingual summarization quality for languages seen in finetuning.

4.4 Impact of Two-Step Finetuning

Row (5) of Table 3 shows our two-step finetuning achieves strong zero-shot results for languages unseen in summarization finetuning. As QK finetuning without the translation step was not capable of cross-lingual zero-shot generation, we have evidence that the model retained knowledge from the crosslingual (translation) training. Also, the two-step finetuning surpasses the pipeline approach on es-ru and tr-tr, where neither the source nor target is English, thereby needing translation twice. This confirms the previous finding on translation error propagation harming summarization quality.

5 Conclusion

In this work, we proposed two methods: 1) QK finetuning and 2) balanced adversarial language classifier to improve intralingual and crosslingual zero-shot summarization. We presented evidence that language-independent representations facilitate zero-shot summarization, in both intralingual and crosslingual forms.

We are curious to see the applicability of our methods to other generation tasks. We are also curious about additional qualitative comparisons of language-specific and -independent representations. In the current study, we only used probing analyses to assess language-specific versus language-independent representations. One way to supplement these analyses is to directly analyze the model hidden representations, e.g., compare the similarity between model hidden representations of different languages before and after applying the proposed approaches. This could for instance be achieved by Singular Vector Canonical Correlation (SVCCA) (Raghu et al., 2017), which has been used to analyze multilingual representations for translation (Kudugunta et al., 2019; Liu et al., 2021; Sun et al., 2023).

Limitations

This works has the following limitations:

Single Underlying Model All out experiments are based on mBART (Liu et al., 2020), specifically the variant pretrained on 25 languages. Extending the current setup to mBART-50 which covers 50 languages can already provide wider language coverage for testing zero-shot inference. Moreover, a further exploration with other pretrained models such as mT5 (Xue et al., 2021) or recent decoder-only large languages models (Scao et al., 2022; Touvron et al., 2023) could further validate the results.

Reliance on Translation Data Our two-step finetuning approach requires many-to-many translation data for the languages of interest. In extremely low-resource cases, we would need to create synthetic data by backtranslation (Sennrich et al., 2016), which requires more computational resources.

Lack of Multiple Experiment Runs Due to computational constraints, the scores in our experiments are reported from single experiment runs. As a partial remedy, we use bootstrap resampling to derive confidence intervals of the reported scores and report the results in Appendix E.

Acknowledgement

We thank the anonymous reviewers for helpful feedback. Part of this work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. Part of this work was supported by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF).

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April*

24-26, 2017, Conference Track Proceedings. OpenReview.net.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7570–7577. AAAI Press.

Nadezhda Chirkova, Sheng Liang, and Vassilina Nikoulina. 2023. Empirical study of pretrained multilingual language models for zero-shot cross-lingual generation. *CoRR*, abs/2310.09917.

Alexandra Chronopoulou, Jonas Pfeiffer, Joshua Maynez, Xinyi Wang, Sebastian Ruder, and Priyanka Agrawal. 2023. Language and task arithmetic with parameter-efficient layers for zero-shot summarization. *CoRR*, abs/2311.09344.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *CoRR*, abs/2308.11466.

Robert M. French and Nick Chater. 2002. Using noise to compute error surfaces in connectionist networks: A novel means of reducing catastrophic forgetting. *Neural Comput.*, 14(7):1755–1769.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen-An Li and Hung-Yi Lee. 2024. Examining forgetting in continual pre-training of aligned large language models. *CoRR*, abs/2401.03129.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 827–838, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1259–1273, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818, Online. Association for Computational Linguistics.

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

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

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmT5: Modular multilingual pre-training solves source language hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1978–2008, Singapore. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6076–6085.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji

Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Haoran Sun, Xiaohu Zhao, Yikun Lei, Shaolin Zhu, and Deyi Xiong. 2023. Towards a deep understanding of multilingual end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14332–14348, Singapore. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Xianze Wu, Zaixiang Zheng, Hao Zhou, and Yong Yu. 2022. LAFT: Cross-lingual transfer for text generation by language-agnostic finetuning. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 260–266, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

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

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Details on Probing Analysis

The probing experiment aims to analyze the model hidden representations regarding the information they contain. Here we are interested in the source language signals in the encoder outputs. We freeze a trained model on the WikiLingua dataset (Ladhak et al., 2020), and train a token-level classifier on the encoder outputs to recover the source language identity, where higher accuracy indicates more language-specific representations. Specifically, we use a linear projection from the hidden dimension to the number of output classes, followed by a softmax activation. For the output classes, we consider the four languages in the crosslingual experiments: English, Spanish, Russian, Turkish. The classifier is trained on the same data as in the summarization task.

B Details on Adversarial Loss

With the standard cross-entropy-based adversarial classifier (Arivazhagan et al., 2019; Mallinson et al., 2020), the classifier itself is trained to optimize

$$\mathcal{L}_{\text{classification}} = -\sum_{c=1}^{L} y_c \log(p_c), \qquad (2)$$

where L is the number of languages, y_c is a binary indicator based on whether the true language label is c, and p_c is the predicted probability for language c. To train the model to deceive the classifier, the adversarial loss is therefore the inverse of Equation 2:

$$\mathcal{L}_{\text{adversarial}} = -\sum_{c=1}^{L} y_c \log(1 - p_c). \qquad (3)$$

However, the term will only be activated when y_c is true, i.e., when the true label is c. This means when the classifier places all its probability mass on another language that is not c (hence still highly language-specific), the adversarial loss will not update the model to change its representations. This leaves the resulting language-specific representations unresolved.

C Dataset Statistics

C.1 Dataset Splits

For the intralingual experiments, we train on XL-Sum (Hasan et al., 2021). Table 4 shows the dataset statistics. For the crosslingual experiments, we train on WikiLingua (Ladhak et al., 2020). Table 5 shows the dataset statistics. For both datasets, in training we exclude samples that have very short inputs (no more than 20 words or punctuation marks) or summaries (no more than 10 words or punctuation marks), as they likely have data quality issues.

	Split	# Samples	Avg. input leng.	Avg. output leng.
English	Train	302,627	459.9	22.3
	Dev	11,535	440.4	21.2
	Test	11,535	437.3	21.2
Spanish	Train	35,633	723.5	29.4
	Dev	4,763	766.5	27.4
	Test	4,763	764.8	27.4
Russian	Train	60,044	564.0	26.1
	Dev	7,780	466.3	24.2
	Test	7,780	465.3	24.2
Gujarati	Train	8,790	769.1	24.0
	Dev	1,139	542.6	21.2
	Test	1,139	529.9	21.7

Table 4: Dataset statistics on XL-Sum. Training is done on English or {English, Spanish, Russian}.

Lang.	Split	# Samples	Lang. pair	Split	# Samples
Intralingual			**Crosslingual**		
en-en	Train	95,517	es-en	Train	76,295
	Dev	3,000		Dev	3,000
	Test	27,489		Test	21,726
es-es	Train	76,295	ru-en	Train	35,313
	Dev	3,000		Dev	3,000
	Test	21,726		Test	9,962
ru-ru	Train	35,313	es-ru	Train	32,458
	Dev	3,000		Dev	3,000
	Test	7,780		Test	8,737
tr-tr	Train	8,790	tr-en	Train	3,052
	Dev	1,139		Dev	438
	Test	1,139		Test	874
			en-tr	Train	3,052
				Dev	438
				Test	874

Table 5: Dataset statistics on WikiLingua. Training is done on intralinuga data in English or {English, Spanish, Russian}.

C.2 Details on Translation Data

We use many-to-many data in all four languages evaluated in the crosslingual experiments: English, Spanish, Russian, and Turkish. To prepare the translation data, we parse the WikiLingua dataset by matching common intputs or outputs of different language pairs. We iterate over samples in different language pairs and match samples that have the same input text or output summary in the same language, but the corresponding output summary or input text is presented in different languages. By performing such matching, we generate translation data in the same domain as used for summarization. A translation model trained with such data is capable of translating both short and long sequences.

D Training and Inference Details

We implement our approaches in the FAIRSEQ (Ott et al., 2019) toolkit.

Training We initialized from the pretrained mBART model[8] (Liu et al., 2020). The word embeddings are frozen due to initial favourable results in zero-shot settings. We use the Adam optimizer (Kingma and Ba, 2015) with betas (0.9, 0.999) and eps 1e-8. We use weight decay of 0.01, start learning rate of 2e-5 and end end learning rate of 5e-9. Dropout is set to 0.1. We use the development set of the same languages as in training for early stopping. All models are trained on an Nvidia Titan

[8]We use the 610M `mbart.CC25` model from https://github.com/facebookresearch/fairseq/blob/main/examples/mbart/README.md#pre-trained-models.

ID	Model / Language	RG-1	RG-2	RG-L	F_{BERT}	L-Acc.
	es-es (en-only train)					
(1)	Full ft.	6.7/6.8/6.9	1.0/1.0/1.1	5.3/5.4/5.5	65.9/66.0/66.1	0.2
(2)	Encoder ft.	24.4/24.8/25.1	6.6/6.9/7.1	18.1/18.4/18.6	70.6/70.8/70.9	85.2
(3)	"LNA" ft.	28.1/28.4/28.8	8.0/8.2/8.5	20.6/20.9/21.1	71.8/71.9/72.1	99.5
(4)	Query-key ft.	**28.3/28.6/29.0**	**8.6/8.8/9.1**	**21.1/21.3/21.7**	**72.1/72.3/72.4**	99.9
(5)	Pipeline	27.8/28.1/28.5	8.0/8.2/8.5	20.4/20.7/21.0	72.0/72.1/72.3	100.0
(6)	Supervised	32.4/32.8/33.3	12.1/12.5/12.9	24.6/25.0/25.4	73.8/74.0/74.2	100.0
	ru-ru (en-only train)					
(1)	Full ft.	1.0/1.0/1.1	0.2/0.2/0.3	0.9/1.0/1.0	64.2/64.3/64.4	2.3
(2)	Encoder ft.	28.0/28.3/28.6	10.2/10.4/10.6	22.4/22.7/22.9	73.1/73.2/73.3	100.0
(3)	"LNA" ft.	26.9/27.3/27.6	9.4/9.6/9.8	21.3/21.6/21.8	72.5/72.7/72.8	100.0
(4)	Query-key ft.	**28.8/29.2/29.5**	**10.9/11.1/11.4**	**23.1/23.4/23.6**	**73.4/73.6/73.7**	100.0
(5)	Pipeline	25.0/25.3/25.6	7.9/8.1/8.3	20.0/20.2/20.5	72.3/72.4/72.5	100.0
(6)	Supervised	33.8/34.1/34.5	14.5/14.7/15.0	27.2/27.5/27.8	74.9/75.1/75.2	100.0
	gu-gu (en-only train)					
(1)	Full ft.	1.2/1.3/1.5	0.2/0.3/0.3	1.1/1.2/1.4	59.0/59.1/59.3	13.4
(2)	Encoder ft.	15.1/15.8/16.6	4.3/4.8/5.3	13.8/14.5/15.2	71.4/71.7/72.0	100.0
(3)	"LNA" ft.	11.2/11.7/12.4	2.8/3.2/3.6	9.9/10.5/11.1	68.3/68.6/68.9	99.8
(4)	Query-key ft.	**17.5/18.3/19.1**	**5.2/5.8/6.4**	**15.9/16.6/17.3**	**72.9/73.2/73.6**	100.0
(5)	Pipeline	14.6/15.2/15.7	2.8/3.2/3.5	13.0/13.6/14.1	71.9/72.1/72.4	100.0
(6)	Supervised	20.8/21.5/22.4	7.1/7.7/8.4	18.5/19.3/20.1	73.9/74.2/74.5	100.0
	gu-gu (multi. train)					
(1)	Full ft.	16.1/16.9/17.6	4.3/4.9/5.4	14.3/15.1/15.8	71.5/71.8/72.0	99.6
(2)	Encoder ft.	16.0/16.8/17.5	4.4/4.9/5.4	14.6/15.3/15.9	71.9/72.2/72.5	100.0
(3)	"LNA" ft.	17.1/17.8/18.5	5.1/5.6/6.1	15.3/16.0/16.8	72.3/72.6/72.8	100.0
(4)	Query-key ft.	**17.4/18.2/19.0**	**5.6/6.2/6.8**	**15.7/16.5/17.3**	**72.8/73.1/73.4**	100.0
(5)	Pipeline	14.6/15.2/15.7	2.8/3.2/3.5	13.0/13.6/14.1	71.9/72.1/72.4	100.0
(6)	Supervised	20.8/21.5/22.4	7.1/7.7/8.4	18.5/19.3/20.1	73.9/74.2/74.5	100.0

Table 6: Full zero-shot intralingual summarization results on XL-Sum calculated using 95% bootstrap confidence intervals (results are presented as 0.025/0.5/0.975 percentiles).

	es-en	ru-en	es-ru	tr-en	en-tr
ROUGE-L	2.2	0.7	0.5	2.3	2.5
L-Acc.	0.0	0.0	0.0	0.0	0.0

Table 7: Results of QK finetuning alone (without two-step finetuning) under the crosslingual zero-shot setup.

RTX GPU with 24GB memory.

Inference When decoding, we use a beam size of 5. The length penalty is 0.6 and 1.0 for intralingual and crosslingual experiments respectively. For the translation model in the pipeline approach, we use the distilled NLLB-200 model (NLLB Team et al., 2022) with 600M parameters.

E Detailed Experiment Scores

Detailed Intralingual Results The detailed results for Table 2 with ROUGE-1 and ROUGE-2 are in Table 6 with RG standing for ROUGE.

Detailed Crosslingual Results The detailed results for Table 3 are in Table 8.

QK Finetuning in Crosslingual Settings QK finetuning alone is not sufficient in crosslingual zero-shot settings. The scores are in Table 7.

ID	Model / Language	RG-1	RG-2	RG-L	F_{BERT}	L-Acc.
	es-en					
(1)	Baseline zero-shot	2.3/2.4/2.4	0.1/0.1/0.1	2.1/2.2/2.2	67.8/67.8/67.9	0.0
(2)	Adv. classifier	33.2/33.4/33.6	10.4/10.5/10.6	26.5/26.7/26.8	76.1/76.1/76.2	98.5
(3)	Balanced adv. classifier	33.9/34.1/34.3	10.8/11.0/11.1	27.1/27.2/27.4	76.3/76.4/76.4	99.4
(4)	(3)+ residual drop	**34.6/34.8/34.9**	11.2/11.3/11.5	27.5/27.6/27.8	**76.5/76.6/76.6**	99.7
(5)	Two-step + QK ft.	34.2/34.4/34.5	**11.3/11.5/11.7**	**27.5/27.7/27.9**	76.4/76.5/76.6	98.4
(6)	Pipeline	37.7/37.9/38.1	14.3/14.5/14.6	31.0/31.1/31.3	78.1/78.1/78.2	99.7
(7)	Supervised	38.2/38.4/38.6	14.6/14.8/14.9	31.2/31.4/31.6	78.1/78.1/78.2	99.8
	ru-en					
(1)	Baseline zero-shot	0.7/0.7/0.7	0.1/0.1/0.1	0.6/0.7/0.7	63.3/63.3/63.3	0.0
(2)	Adv. classifier	31.6/31.9/32.2	9.7/9.9/10.1	25.1/25.3/25.5	75.6/75.7/75.7	99.6
(3)	Balanced adv. classifier	31.9/32.2/32.5	10.0/10.2/10.4	25.4/25.6/25.9	75.7/75.8/75.9	99.8
(4)	(3)+ residual drop	**32.9/33.1/33.4**	10.4/10.6/10.8	**26.0/26.3/26.5**	**76.0/76.1/76.2**	99.9
(5)	Two-step + QK ft.	32.6/32.8/32.9	**10.6/10.7/10.8**	**26.1/26.3/26.4**	**76.0/76.1/76.3**	99.6
(6)	Pipeline	34.4/34.6/34.9	12.0/12.2/12.4	28.2/28.5/28.7	77.2/77.3/77.4	99.7
(7)	Supervised	35.7/36.0/36.3	12.9/13.2/13.4	29.2/29.4/29.7	77.4/77.5/77.5	99.7
	es-ru					
(1)	Baseline zero-shot	0.5/0.6/0.6	0.1/0.1/0.1	0.5/0.6/0.6	64.6/64.6/64.7	0.0
(2)	Adv. classifier	16.6/16.9/17.1	3.9/4.1/4.2	13.8/14.1/14.3	72.4/72.5/72.6	97.6
(3)	Balanced adv. classifier	17.1/17.3/17.5	4.1/4.3/4.4	14.1/14.3/14.5	72.7/72.8/72.9	99.9
(4)	(3)+ residual drop	17.6/17.8/18.0	4.3/4.5/4.6	**14.5/14.8/15.0**	73.0/73.1/73.2	100.0
(5)	Two-step + QK ft.	**17.4/17.6/17.8**	**4.5/4.6/4.6**	**14.7/14.8/14.9**	**73.2/73.4/73.6**	98.4
(6)	Pipeline	16.4/16.7/16.9	3.5/3.7/3.8	14.2/14.4/14.6	73.7/73.8/73.8	100.0
(7)	Supervised	20.8/21.0/21.3	6.0/6.1/6.3	17.7/18.0/18.2	75.1/75.2/75.3	100.0
	tr-en					
(1)	Baseline zero-shot	4.4/5.0/5.6	0.9/1.1/1.4	4.1/4.6/5.1	62.6/62.9/63.1	1.6
(2)	Adv. classifier	31.0/32.0/33.0	10.0/10.7/11.4	25.3/26.1/26.9	74.9/75.2/75.5	98.9
(3)	Balanced adv. classifier	31.7/32.6/33.6	10.3/11.0/11.7	25.7/26.6/27.4	75.2/75.5/75.8	99.8
(4)	(3)+ residual drop	31.2/32.1/33.0	9.7/10.3/11.0	24.9/25.7/26.5	74.9/75.2/75.5	99.1
(5)	Two-step + QK ft.	**38.3/38.6/38.8**	**14.4/14.6/14.7**	**30.5/30.7/30.9**	**77.1/77.4/77.6**	99.7
(6)	Pipeline	39.9/40.9/41.8	16.1/17.0/17.8	33.2/34.1/35.0	78.4/78.7/79.0	99.4
(7)	Supervised	40.4/41.4/42.5	17.1/18.1/19.1	33.5/34.5/35.5	78.5/78.8/79.1	99.4
	en-tr					
(1)	Baseline zero-shot	2.4/2.7/3.0	0.4/0.5/0.5	2.3/2.5/2.7	60.8/60.9/61.1	0.0
(2)	Adv. classifier	2.5/2.8/3.0	0.4/0.5/0.6	2.3/2.5/2.8	60.7/60.9/61.0	0.0
(3)	Balanced adv. classifier	2.5/2.8/3.1	0.4/0.5/0.6	2.4/2.6/2.8	60.8/60.9/61.1	0.0
(4)	(3)+ residual drop	2.4/2.7/3.0	0.4/0.5/0.6	2.3/2.5/2.7	60.8/61.0/61.1	0.0
(5)	Two-step + QK ft.	**20.2/20.4/20.6**	**5.7/5.8/5.8**	**16.5/16.7/16.9**	**71.0/71.3/71.5**	98.7
(6)	Pipeline	20.1/20.9/21.6	5.1/5.5/6.0	18.0/18.7/19.4	72.8/73.1/73.4	99.8
(7)	Supervised	22.7/23.7/24.8	7.4/8.0/8.7	19.7/20.7/21.5	72.9/73.2/73.5	100.0
	tr-tr					
(1)	Baseline zero-shot	20.0/20.9/21.7	5.5/6.0/6.5	17.4/18.0/18.8	71.2/71.5/71.8	96.4
(2)	Adv. classifier	5.3/5.7/6.2	1.0/1.1/1.3	4.7/5.2/5.6	62.6/62.8/63.0	10.6
(3)	Balanced adv. classifier	3.2/3.5/3.8	0.5/0.5/0.7	3.0/3.2/3.5	60.9/61.1/61.3	0.1
(4)	(3)+ residual drop	2.2/2.5/2.7	0.3/0.4/0.5	2.1/2.3/2.5	60.6/60.8/60.9	0.0
(5)	Two-step + QK ft.	**22.9/23.1/23.3**	**6.9/7.0/7.1**	**18.2/18.4/18.6**	**71.7/72.0/72.2**	100.0
(6)	Pipeline	19.9/20.7/21.5	5.4/5.8/6.3	17.7/18.5/19.2	73.0/73.2/73.5	99.8
(7)	Supervised	29.2/30.3/31.3	11.4/12.3/13.0	25.3/26.2/27.2	75.2/75.4/75.7	99.7

Table 8: Full zero-shot crosslingual summarization results on WikiLingua calculated using 95% bootstrap confidence intervals (results are presented as 0.025/0.5/0.975 percentiles).