Exploring the Impact of Layer Normalization for Zero-shot Neural Machine Translation

Zhuoyuan Mao [1]   **Raj Dabre** [2]   **Qianying Liu** [1]
Haiyue Song [1]   **Chenhui Chu** [1]   **Sadao Kurohashi** [1,3]
[1] Kyoto University, Japan   [2] NICT, Japan   [3] NII, Japan
{zhuoyuanmao, ying, song, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp
raj.dabre@nict.go.jp

Abstract

This paper studies the impact of layer normalization (LayerNorm) on zero-shot translation (ZST). Recent efforts for ZST often utilize the Transformer architecture as the backbone, with LayerNorm at the input of layers (PreNorm) set as the default. However, Xu et al. (2019) has revealed that PreNorm carries the risk of overfitting the training data. Based on this, we hypothesize that PreNorm may overfit supervised directions and thus have low generalizability for ZST. Through experiments on OPUS, IWSLT, and Europarl datasets for 54 ZST directions, we demonstrate that the original Transformer setting of LayerNorm after residual connections (PostNorm) consistently outperforms PreNorm by up to 12.3 BLEU points. We then study the performance disparities by analyzing the differences in off-target rates and structural variations between PreNorm and PostNorm. This study highlights the need for careful consideration of the LayerNorm setting for ZST.

1 Introduction

Multilingual neural machine translation (MNMT) enables translation between unseen language pairs, i.e., zero-shot translation (ZST) (Johnson et al., 2017; Firat et al., 2017). Prior studies have explored techniques such as language tags (Wu et al., 2021), residual connections (Liu et al., 2021), and novel training objectives (Al-Shedivat and Parikh, 2019; Pham et al., 2019; Arivazhagan et al., 2019; Gu et al., 2019; Zhu et al., 2020; Zhang et al., 2020; Wang et al., 2021; Yang et al., 2021) for improving ZST. They primarily used the Transformer architecture (Vaswani et al., 2017), which has two variations depending on the position of layer normalization (LayerNorm) (Ba et al., 2016), namely, PreNorm (applied at the input of layers) (Baevski and Auli, 2019) and PostNorm (applied after residual connections), as shown in Fig. 1. As previous studies showed that PreNorm can result in more stable training and faster convergence compared to
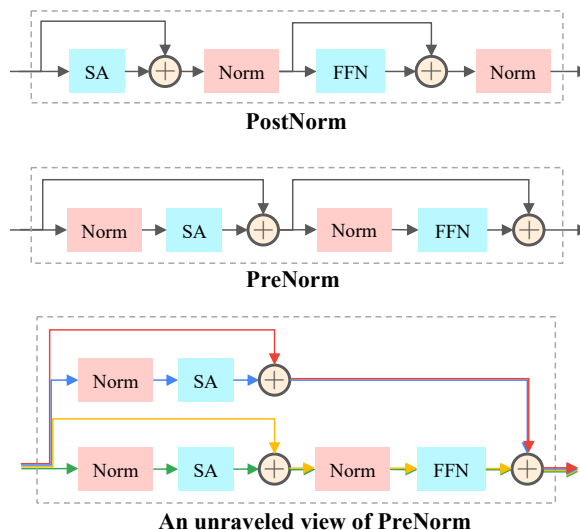

Figure 1: **PostNorm, PreNorm, and an unraveled view of PreNorm in a Transformer encoder layer**. "Norm," "SA," and "FFN" denote LayerNorm, self-attention, and feed-forward network. $\oplus$ is residual connection. Paths with different colors in the unraveled view of PreNorm indicate respective sub-networks.

PostNorm for MNMT (Xiong et al., 2020), most ZST works (Pham et al., 2019; Wu et al., 2021; Liu et al., 2021) use PreNorm as the default setting following those MNMT studies. However, Xu et al. (2019) revealed that PreNorm carries the risk of overfitting the training data. We thus hypothesize that in a multilingual scenario, PreNorm may overfit supervised directions and have poor ZST generalizability. We systematically explore PreNorm and PostNorm's effect on ZST to verify this.

Using the OPUS, IWSLT, and Europarl datasets and a total of 54 ZST directions, we show that PostNorm consistently outperforms PreNorm by up to 12.3 BLEU points. Following previous work, we also evaluate different language tag (Wu et al., 2021) and residual connection (Liu et al., 2021) settings, as they have been shown to impact ZST but we observe that PostNorm continues to be superior thereby lending credibility to our hypothesis.

1300

To better understand the performance differences, we introduce a novel analysis approach called **layer-wise language recognition** (**LLR**), which tracks the off-target rates for each encoder and decoder layer by training token-level classifiers to recognize the source or target language. This analysis shows that PreNorm is more sensitive to language tag settings than PostNorm, negatively impacting ZST performance. Additionally, by examining the unraveled view of PreNorm (Fig. 1) inspired by Veit et al. (2016), we reveal structural flaws in PreNorm for ZST. Our analysis demonstrates that the order of LayerNorm and self-attention/feed-forward network in PreNorm is the main factor affecting its ZST performance.

Given the prevalent use of PreNorm as the default setting in ZST baselines and frameworks such as Fairseq (Ott et al., 2019)[1] and Tensor2Tensor (Vaswani et al., 2018), our study emphasizes the importance of careful consideration in the LayerNorm setting for ZST.

2 Background: LayerNorm

LayerNorm (Ba et al., 2016) normalizes the input **x** by zero-centering and scaling to have a unit standard deviation, followed by an additional trainable transformation, including a gain and bias adjustment. Specifically, it is formulated as:

$$\text{LayerNorm}(\mathbf{x}) = \frac{\mathbf{x} - \mathbf{E}(\mathbf{x})}{\sqrt{\mathbf{V}(\mathbf{x})}} \cdot \mathbf{g} + \mathbf{b}, \quad (1)$$

where **g** and **b** are trainable gain and bias. **E** and **V** indicate expectation and variance. LayerNorm is commonly used in two positions in the Transformer, as shown in Fig. 1. PostNorm, which is the originally proposed setting of the Transformer (Vaswani et al., 2017), involves applying LayerNorm after each sub-module (i.e., self-attention or feed-forward network) and residual connections. PreNorm (Baevski and Auli, 2019), on the other hand, involves applying LayerNorm directly before each sub-module and is known to stabilize Transformer training. While variants of Transformer LayerNorm like RMSNorm (Zhang and Sennrich, 2019) have been proposed, the vanilla PreNorm and PostNorm are still the most widely adopted settings in current multilingual NMT literature. Therefore, we only focus on PreNorm and PostNorm in this work.

Datasets	Languages	$\mathbf{N}_{zero}$	$\mathbf{S}_{train}$	Arch.
OPUS	ar, de, en, fr, nl, ru, zh	30	12.00M	base
IWSLT	en, it, nl, ro	6	1.38M	base
Europarl	de, en, es, fr, nl	12	15.78M	big

Table 1: **Statistics of the training data**. $\mathbf{N}_{zero}$ and $\mathbf{S}_{train}$ denote number of the ZST directions and size of the training data, respectively. `base` and `big` indicate `Transformer-base` and `Transformer-big`.

Nguyen and Salazar (2019) have explored the impacts of normalization and initialization choices on supervised low-resource NMT settings, however, we delve deeper and focus on the significance of the positioning of LayerNorm for zero-shot NMT. We expect this to complete the understanding of LayerNorm's role in multilingualism, particularly in the context of zero-shot translation.

3 Experiments and Results

We evaluate the performance of PreNorm and PostNorm for ZST on various datasets and language pairs. We then analyze the off-target rates and structural discrepancies between PreNorm and PostNorm to understand performance differences.

3.1 Experimental Settings

Datasets We perform ZST experiments on three datasets: OPUS (Zhang et al., 2020), IWSLT (Cettolo et al., 2017), and Europarl (Koehn, 2005). The statistics of the datasets are summarized in Table 1. We include 7, 4, and 5 languages for each dataset. The training data consists of only English-centric sentence pairs, resulting in 30, 6, and 12 ZST directions for each dataset. The total number of parallel sentences for each dataset is 12.00M, 1.38M, and 15.78M, respectively. We apply BPE (Sennrich et al., 2016) with merge operations of 50k, 40k, and 50k to create a joint vocabulary for each dataset.

Training We employ `Transformer-base` model for OPUS and IWSLT, and `Transformer-big` for Europarl, in accordance with the distinct sizes of training data. We consider the following settings:
(1) PreNorm or PostNorm: PreNorm involves LayerNorm directly before each sub-module (i.e., self-attention or feed-forward network), while PostNorm applies LayerNorm after each sub-module and residual connections, as shown in Fig. 1.[2]

#	Layer Norm	Language Tag	Res.	Zero-shot			Supervised		
				OPUS	IWSLT	Europarl	OPUS	IWSLT	Europarl
0		*Pivot*		21.8	20.0	29.5	-	-	-
1	**PreNorm**	S-ENC-T-DEC	w/	10.1 (42.19%)	4.9 (64.84%)	24.9 (7.73%)	33.7	31.5	34.3
2	**PostNorm**	S-ENC-T-DEC	w/	**16.8** (**8.59%**)	**12.4** (**10.61%**)	**29.2** (**0.34%**)	33.9	31.5	34.5
3	**PreNorm**	T-ENC	w/	13.3 (22.99%)	13.7 (3.98%)	29.5 (0.23%)	33.7	31.6	34.4
4	**PostNorm**	T-ENC	w/	**14.0** (**22.86%**)	**15.5** (4.59%)	**30.8** (**0.11%**)	34.1	31.5	34.5
5	**PreNorm**	S-ENC-T-DEC	w/o	14.3 (20.67%)	8.0 (50.16%)	16.7 (41.87%)	33.6	30.9	34.3
6	**PostNorm**	S-ENC-T-DEC	w/o	**16.0** (**15.27%**)	**17.4** (**1.83%**)	**29.0** (**0.41%**)	33.8	30.7	34.4
7	**PreNorm**	T-ENC	w/o	13.4 (27.15%)	16.2 (1.54%)	29.9 (2.15%)	33.5	30.9	34.3
8	**PostNorm**	T-ENC	w/o	**13.9** (**26.68%**)	**17.8** (**1.50%**)	**30.8** (**0.13%**)	33.9	30.6	34.4

Table 2: **BLEU scores and off-target rates (shown in brackets)**. We report the average score of three seeds; refer to Appendix G for BLEU score of each translation direction and seed. "Res." indicates the residual connection of self-attention in the 4th encoder layer. We mark lower off-target rates and significantly higher BLEU scores (Koehn, 2004) between PreNorm and PostNorm in **bold** for ZST.

(2) S-ENC-T-DEC or T-ENC: Source language tag on the encoder-side and target language tag on the decoder-side; or only target language tag on the encoder-side. Wu et al. (2021) showed that this setting impacts ZST for Transformer with PreNorm.
(3) w/ or w/o Res.: With the residual connection for self-attention in the middle (4th) encoder layer or not. Liu et al. (2021) revealed that "w/o Res." improves ZST for the model trained with PreNorm. We experiment this with different LayerNorm settings as this may reduce the potential of overfitting on supervised directions, then further impacts ZST, which aligns with our hypothesis.

The settings above lead to eight different combinations, shown in Table 2 (#1 - #8). Additional training details are in Appendix A.

3.2 Main Results

We evaluate ZST systems using SacreBLEU (Post, 2018) and off-target rates. We report in Table 2 BLEU scores for both zero-shot and supervised directions. For ZST, we also present pivot-based translation results as a reference. Implementation details of evaluation can be found in Appendix B. Our findings are as follows:
PreNorm vs. PostNorm: We find that Post-Norm consistently yields better BLEU scores than PreNorm for ZST across various language tag and residual connection settings, while their performance is comparable for supervised directions.
Impact of Language Tag and Residual Connection: We observe that using the "T-ENC" language tag and "w/ Res." improves ZST performance for IWSLT, which aligns with the findings of Wu et al. (2021) and Liu et al. (2021). Nevertheless, the best performance is achieved using "w/ Res." for Post-

Norm with "S-ENC-T-DEC" and "T-ENC" tags for OPUS and Europarl, respectively (#2 and #4). Given that Wu et al. (2021) and Liu et al. (2021) used PreNorm as the default setting (#2, #4, #6 and #8 are unreported results in their work), our results emphasize the need to consider PostNorm as the default setting for ZST, while the language tag and residual connection settings have less impact.
Off-target Rates: Off-target rates help understand the different BLEU score gaps between PreNorm and PostNorm, which ranges from 0.5 to 12.3 BLEU points. For PreNorm and PostNorm with the "T-ENC" language tag (#3, #4, #7, and #8), they have similar off-target rates, with a discrepancy ranging from -0.61% to 2.02%, which results in narrow BLEU score gaps, ranging from 0.5 to 1.8 points. However, for PreNorm and PostNorm with the "S-ENC-T-DEC" language tag (#1, #2, #5, and #6), the off-target rates show a more considerable discrepancy, ranging from 5.40% to 54.23%, resulting in BLEU score gaps from 1.7 to 12.3 points. Further analysis of the nature of Transformer hidden states in the next section explores the reason for these different off-target rates in translations.

3.3 Tracking Off-targets within Transformer

We probe the language independence of hidden states to track off-targets within Transformer and reveal the differences between PreNorm and PostNorm. In previous work, language independence was primarily analyzed using either SVCCA (Raghu et al., 2017) or language classification accuracy (LCA) (Liu et al., 2021). However, we provide evidence in Appendix C that SVCCA, which measures the cosine similarity between hidden states, are not suitable for ZST systems. In-
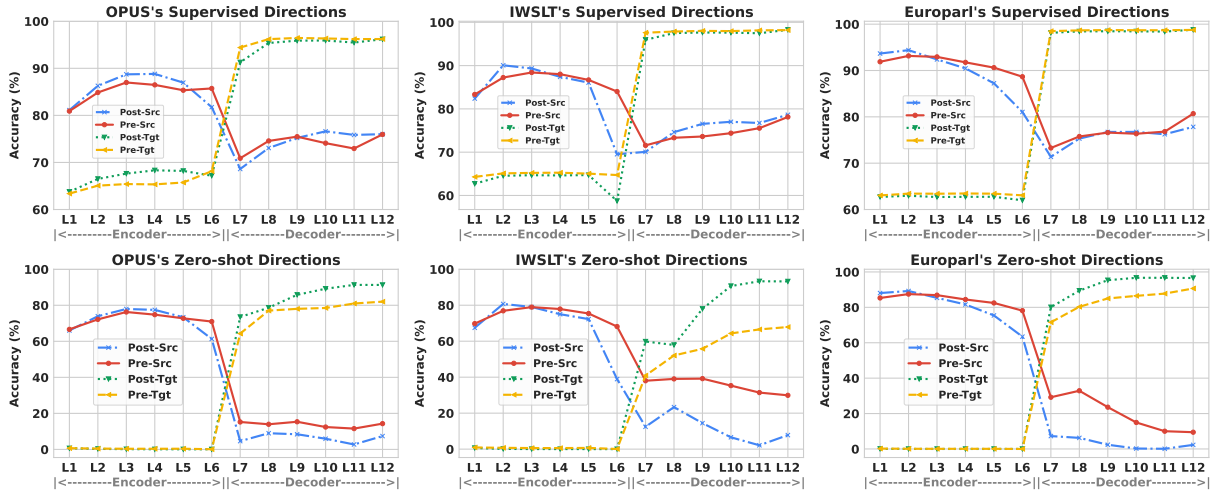
Figure 2: **The LLR results of #1 and #2 (Table 2) for both ZST and supervised directions for each dataset**. We report the average accuracy of three seeds and all the supervised or zero-shot directions. "Pre-Src" and "Pre-Tgt" indicate the layer-wise source and target language recognition for a PreNorm system (#1), while "Post-Src" and "Post-Tgt" denote similary for a PostNorm system (#2). "L1" to "L6" are 6 encoder layers and "L7" to "L12" are 6 decoder layers. We present the figures of other systems (#3 - #8) in Appendix F.

stead, LCA trains a classifier to inspect the hidden states on top of the encoder, but it does not simulate the training of a ZST system, which may introduce bias in the analysis for ZST.[3] In this work, we propose a novel approach for ZST based on LCA:

LLR tailors classifiers for each layer to recognize the source or target language. We train a token-level linear classifier for each layer to utilize hidden states in each layer as features to identify the source or target language. We use hidden states obtained by feeding sentence pairs in supervised directions to simulate the training of ZST. We then test each layer's classifer's ability to recognize the source or target language for supervised or zero-shot directions. This approach enables the trained classifier to best represent the language recognition ability of hidden states in a ZST system.

We train two types of linear classifiers for each encoder and decoder layer. One is for recognizing the source language, and the other is for the target language. Each linear classifier is a linear transformation from the dimension of the hidden states (512 or $1,024$) to the number of source or target languages (e.g., 7 for OPUS). We use the validation set of all supervised directions to obtain the hidden

state of each token in each layer and set their source language tag or target language tag as the gold labels. Note that the decoder hidden state of each token in each layer is obtained auto-regressively without teacher-forcing. We train each classifier for 3 epochs[4] with a learning rate of 1e-3 and a batch size of 64 sentences. For inference, we utilize the test sets of all supervised or zero-shot directions for computing the LLR results for corresponding directions, respectively.

The LLR results for #1 and #2 in Table 2 are presented in Fig. 2. First, we find that the encoder and decoder hidden states are highly correlated with the target and source languages, respectively, for supervised directions (L1 to L6 of Pre/Post-Tgt and L7 to L12 of Pre/Post-Src of 3 upper sub-figures), which may impact the generalizability for ZST. Second, we see that the encoder hidden states of Post-Norm are less dependent on the source language than PreNorm (L6 of Pre/Post-Src of 3 lower sub-figures). Third, we observe that the hidden states in all the decoder layers of PostNorm are more dependent on the target language and less on the source language than PreNorm (L7 to L12 of 3 lower sub-figures). The latter two points contribute to the observed gaps in off-target rates between PreNorm and PostNorm. Conclusions for #5 and #6 with the "S-ENC-T-DEC" tag are identical (Appendix G).

---

[3]Liu et al. (2021) regulate the output language via a decoder-side language tag, hence analyzing only the encoder states poses no issues as the target language tag does not impact them. Nevertheless, with other language tag settings such as S-ENC-T-DEC and T-ENC, employed in this study, we require a method to obtain hidden states properly, given their impact on hidden states.

[4]The classifier can fully converge within 3 epochs as the classifier is lightweight that only contains a small number of parameters.
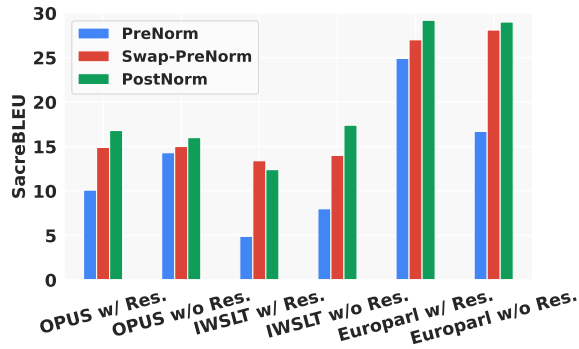
Figure 3: **BLEU scores of systems with "S-ENC-T-DEC" for ZST**. We report the mean of three seeds.

For systems using "T-ENC," we find that the LLR are similar between PreNorm and PostNorm (Appendix G) and attribute the BLEU score gaps to translation quality (i.e., adequacy and fluency).

3.4 Unraveling Structural Flaws of PreNorm

We investigate the structural differences between PreNorm and PostNorm to explain the observed differences in hidden states for models trained with the "S-ENC-T-DEC" tag. Inspired by Veit et al. (2016), we present an "unraveled view" for PreNorm, which decomposes the residual connections by the summation of several sub-networks, as shown in Fig. 1 (paths with different colors indicate sub-networks). However, this is not applicable to PostNorm, as LayerNorm is located after residual connections. Based on this analysis, the structural characteristic of PreNorm is:

(1) Shallow Sub-network Nature: PreNorm includes shallow sub-networks, such as the embedding layer output fed through encoder layers without any operation except for the final LayerNorm (red path in Fig. 1), but PostNorm does not.

(2) LayerNorm Before SA/FFN: In PreNorm, LayerNorm is placed directly before the self-attention (SA) or feed-forward module (FFN) within the residual connection module.

To analyze the impact of these structural characteristics on the generalizability of PreNorm in ZST, we swap the order of LayerNorm and SA/FFN within the residual connection module (**Swap-PreNorm**), while keeping the shallow sub-network nature of PreNorm. Refer to Appendix D for specific illustrations of Swap-PreNorm. The results, presented in Fig 3, show that PreNorm can be significantly improved through Swap-PreNorm, with Swap-PreNorm approaching the performance of PostNorm. This demonstrates that ZST is more

sensitive to the position of LayerNorm in PreNorm than its shallow sub-network nature.

4 Conclusion

In this paper, we comprehensively explored the effects of LayerNorm on ZST performance. Our results demonstrate that PostNorm consistently outperforms PreNorm for ZST, regardless of the language tag and residual connection settings used. Through in-depth analysis of off-target rates and structural flaws in the PreNorm model, we were able to identify the underlying factors that contribute to the performance discrepancy. Our study suggests that care should be taken when selecting the LayerNorm setting for ZST in future research.

Limitations

According to us there are 3 limitations of our work which will be addressed in future work.

- The impact of LayerNorm, language tags, and residual connection settings on ZST was analyzed in this study. However, other factors, such as the number of layers of the Transformer model, may also have an effect and should be further investigated.

- Our conclusions were based on overall scores across all ZST directions. Further examination of how LayerNorm impacts specific language pairs is necessary.

- We explored the setting of LayerNorm for ZST systems trained from scratch. Exploration of how the LayerNorm setting of multilingual pre-trained models such as mBART (Liu et al., 2020) impacts the fine-tuning for ZST will be needed.

Ethical Considerations

In this study, we utilized only publicly accessible datasets for model training. Though our experiments focused on neural machine translation models, it is worth noting that these models may produce biased translations. Although this can be mitigated through a debiasing filtering process, it is beyond the scope of this work. Regarding the composition of this paper, only Grammarly[5] was utilized for grammar correction, and there is no originally machine-generated text in the paper.

---

[5]https://app.grammarly.com/

Acknowledgements

References

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman-Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Comput. Speech Lang.*, 45:236–252.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. 2016. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

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*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. 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.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

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.

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.

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

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Ashish Vaswani, Samy Bengio, Eugene Brevdo, François Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 193–199. Association for Machine Translation in the Americas.

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

Andreas Veit, Michael J. Wilber, and Serge J. Belongie. 2016. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 550–558.

Weizhi Wang, Zhirui Zhang, Yichao Du, Boxing Chen, Jun Xie, and Weihua Luo. 2021. Rethinking zero-shot neural machine translation: From a perspective of latent variables. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4321–4327, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. Understanding and improving layer normalization. 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 4383–4393.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. Improving multilingual translation by representation and gradient regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7266–7279, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. 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 12360–12371.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. 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.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1650–1655, Online. Association for Computational Linguistics.

A Training Details

For data preprocessing, we utilize jieba[6] for Chinese segmentation and Moses[7] (Koehn et al., 2007) for tokenization of other languages. After applying BPE, we obtain vocabularies with sizes of $66, 158$, $40, 100$, and $50, 363$ for OPUS, IWSLT, and Europarl, respectively. For multilingual training, we do not apply oversampling as the data size for each language pair is comparable. The maximum sentence length is set to 256. We train Transformer models using Fairseq[8] and set the dropout rate to 0.1, 0.4, and 0.3 for each dataset. Adam (Kingma and Ba, 2015) is used as the optimizer with a learning rate of 5e-4, 1e-3, and 5e-4 for each dataset, and $4, 000$ warm-up steps are employed. We train the Transformer-base model using 4 32G V100 GPUs and the Transformer-big model using 8 32G V100 GPUs with the batch size of $4, 096$ tokens. Additionally, we employ mixed precision training (Micikevicius et al., 2018) to accelerate the training process. We train each dataset for 200, 100, and 400 epochs, respectively.

B Evaluation Details

For OPUS, we use the test sets following (Zhang et al., 2020), while for IWSLT and Europarl, we choose the test sets following (Wu et al., 2021). We select the checkpoint with the lowest validation loss for evaluation. The inference is performed on the trained models using a beam size of 5. For calculating SacreBLEU,[9] we utilize the "zh" tokenization mode for Chinese, and the "13a" tokenization mode for other languages. We use the model of setting #4[10] (Table 2) for pivot-based translation. To calculate the off-target rates, we utilize the language identification tool provided by FastText (Joulin et al., 2016).[11] Our experiment has revealed that this tool is slightly more accurate than another tool called "langdetect,"[12] as it can achieve

---

[6] https://github.com/fxsjy/jieba
[7] https://github.com/moses-smt/mosesdecoder
[8] https://github.com/facebookresearch/fairseq
[9] https://github.com/mjpost/sacrebleu
[10] We use this setting as it achieves the best performance for supervised directions, as shown in Table 2.
[11] https://fasttext.cc/docs/en/language-identification.html
[12] https://github.com/Mimino666/langdetect

	Zero-shot	Supervised
PreNorm	9.8	33.8
PostNorm	17.5	33.8
PreNorm w/o Enc-Last	11.2	33.7

Table 3: **BLEU scores of PreNorm, PostNorm, and "PreNorm w/o Enc-Last" on OPUS**. They are trained with the "S-ENC-T-DEC" tag, "Res.," and the random seed of 10. We report the mean of all the translation directions.
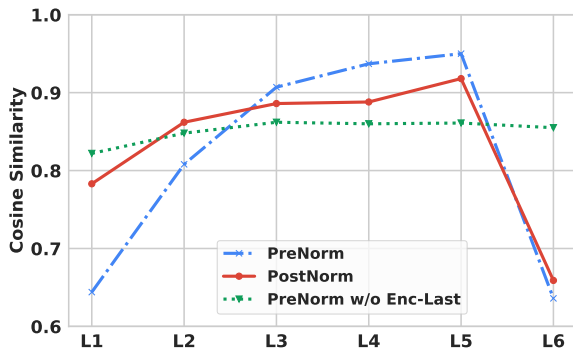
Figure 4: **Encoder layer-wise SVCCA scores of PreNorm, PostNorm, and "PreNorm w/o Enc-Last" between "en-xx" and "xx-en" translation directions**. We report the mean of all the direction pairs.

an accuracy of 98% when decoding reference English sentences in the test set, whereas "langdetect" only achieves accuracy of around 92%.

C Discussion about SVCCA score

In previous work (Wu et al., 2021; Liu et al., 2021), the SVCCA score (Raghu et al., 2017), a cosine similarity measure between the hidden states of neural models, was used to compare two ZST models. However, we demonstrate that this method is unsuitable for comparing different ZST systems through an experiment. We removed the final LayerNorm from the PreNorm encoder, denoting it as "PreNorm w/o Enc-Last." We then evaluated the BLEU scores of PreNorm, PostNorm, and "PreNorm w/o Enc-Last" on the OPUS dataset, as reported in Table 3. We subsequently calculated the encoder layer-wise SVCCA score for each LayerNorm setting using the mean-pooled hidden states of each encoder layer. The average SVCCA score between all the "en-xx" and "xx-en" directions is reported in Fig. 4. When comparing Fig. 4 with Table 3, we observe that PostNorm has a higher SVCCA score on top of the encoder (L6) than PreNorm, which suggests that the encoder of Post-
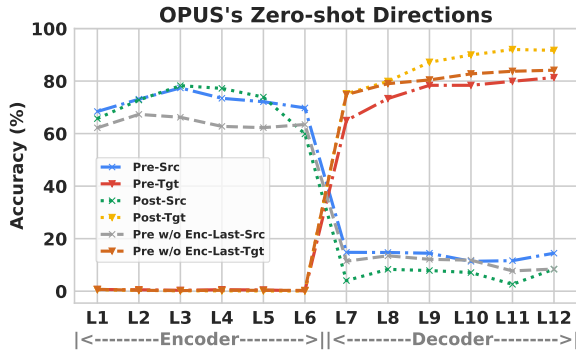
Figure 5: **The LLR results of PreNorm, PostNorm, and "PreNorm w/o Enc-Last."** We report the mean of all the ZST directions. "-Src" and "-Tgt" indicate the LLR results for the source and target languages, respectively. "L1" to "L6" are 6 encoder layers and "L7" to "L12" are 6 decoder layers.

Figure 6: **Swap-PreNorm, and an unraveled view of Swap-PreNorm in a Transformer encoder layer**. "Norm," "SA," and "FFN" denote LayerNorm, self-attention, and feed-forward network. ⊕ is residual connection. Paths with different colors in the unraveled view of PreNorm indicate respective sub-networks.

Norm is more language-agnostic and thus has a higher ZST BLEU score in Table 3, aligning with the results found in Wu et al. (2021) and Liu et al. (2021). However, "PreNorm w/o Enc-Last" shows an extremely high SVCCA score on top of the encoder, whereas its ZST BLEU performance is significantly lower than PostNorm by 6.3 BLEU points. This reveals the significant inconsistency between the SVCCA score and the performance of ZST models. Therefore, it is crucial to carefully consider how to leverage SVCCA for ZST analysis in the future.

On the other hand, our proposed LLR score is consistent with the ZST BLEU score, as shown in Fig. 5. Specifically, we observe the lowest LLR score on top of the encoder of PostNorm for the source language and the highest LLR scores in all the decoder layers, which aligns with its best ZST performance among the three systems.

D Swap-PreNorm

Fig. 6 illustrates the implementation of Swap-PreNorm, which incorporates LayerNorm following the SA/FFN layers within the residual connection block. Compared with PostNorm, Swap-PreNorm alters the order of LayerNorm and residual connections. As depicted in the unraveled view of Swap-PreNorm in Fig. 6, it preserves the shallow sub-network characteristics of PreNorm, which is the main difference compared with PostNorm.

E LayerNorm without Trainable Parameters

Xu et al. (2019) demonstrated that the overfitting issue of PreNorm can be alleviated by removing the trainable parameters of LayerNorm (LayerNorm-simple). We apply this technique to our ZST experimental settings to investigate the overfitting state of PreNorm and PostNorm. PreNorm and Post-Norm after applying this technique are denoted as PreNorm-simple and PostNorm-simple. As reported in Table 4, the results indicate that PreNorm-simple and PostNorm-simple outperform their respective original versions in supervised directions, which aligns with the findings of Xu et al. (2019). Additionally, we observe comparable or better BLEU scores for PreNorm-simple than PreNorm (except for #7 on Europarl), indicating that the original PreNorm had low generalizability for ZST. For PostNorm-simple, we observe significant improvement only for #4 on OPUS, which suggests the superior generalizability of the original PostNorm for ZST. Despite this improvement, PreNorm-simple still underperforms PostNorm, highlighting the severe overfitting problem of the original PreNorm.

F Details of the LLR Results

We show the LLR results of #3 - #8 (Table 2) for ZST and supervised directions in Fig. 7.

G Details of the Main Results

We report the specific BLEU score for each translation direction and each random seed in Ta-

#	LayerNorm-simple	Language Tag	Res.	Zero-shot			Supervised		
				OPUS	IWSLT	Europarl	OPUS	IWSLT	Europarl
1	**PreNorm-simple**	S-ENC-T-DEC	w/	10.1 (+0.0)	5.9 (+1.0)	25.0 (+0.1)	33.9 (+0.2)	31.9 (+0.4)	34.4 (+0.1)
2	**PostNorm-simple**	S-ENC-T-DEC	w/	**15.8** (-1.0)	**11.5** (-0.9)	**28.7** (-0.5)	34.1 (+0.2)	32.1 (+0.6)	34.5 (+0.0)
3	**PreNorm-simple**	T-ENC	w/	13.7 (+0.4)	14.5 (+0.8)	29.4 (-0.1)	33.9 (+0.2)	31.9 (+0.3)	34.4 (+0.0)
4	**PostNorm-simple**	T-ENC	w/	**14.9** (+0.9)	**15.4** (-0.1)	**30.8** (+0.0)	34.0 (-0.1)	31.9 (+0.4)	34.6 (+0.1)
5	**PreNorm-simple**	S-ENC-T-DEC	w/o	15.4 (+1.1)	7.8 (-0.2)	19.4 (+2.7)	33.7 (+0.1)	31.3 (+0.4)	34.1 (-0.2)
6	**PostNorm-simple**	S-ENC-T-DEC	w/o	**16.4** (+0.4)	**16.0** (-1.4)	**29.2** (+0.2)	33.9 (+0.1)	31.3 (+0.6)	34.4 (+0.0)
7	**PreNorm-simple**	T-ENC	w/o	13.1 (-0.3)	16.8 (+0.6)	28.7 (-1.2)	33.7 (+0.2)	31.4 (+0.5)	34.3 (+0.0)
8	**PostNorm-simple**	T-ENC	w/o	**14.0** (+0.1)	**17.9** (+0.1)	**31.0** (+0.2)	33.7 (-0.2)	31.1 (+0.5)	34.4 (+0.0)

Table 4: **BLEU scores of LayerNorm-simple**. We report the average score of three seeds. "Res." indicates the residual connection of self-attention in the 4th encoder layer. We mark better scores between PreNorm-simple and PostNorm-simple in **bold**. For each setting, significantly better or worse BLEU scores (Koehn, 2004) compared with the results in Table 2 are marked in blue or red.

bles 5, 6, 7, 8, 9, and 10.[13] In addition to BLEU scores, we present model-based evaluation results obtained using BLEURT (Sellam et al., 2020)[14] in Table 11. The results trend is consistent with those obtained from BLEU scores.

---

[13] Refer to details of setting random seeds in PyTorch at https://pytorch.org/docs/stable/notes/randomness.html.

[14] https://github.com/google-research/bleurt
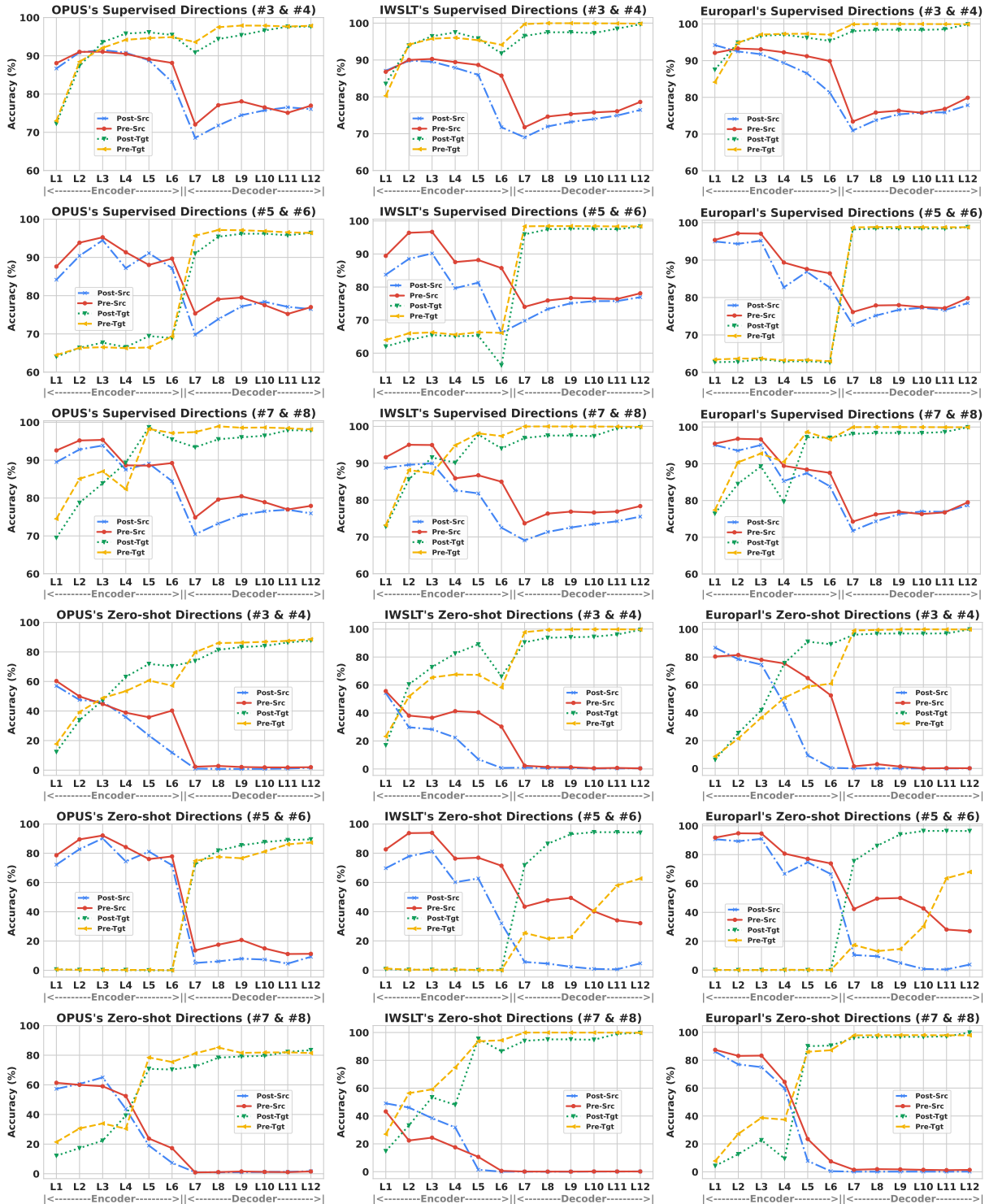
Figure 7: **The LLR results of #3 - #8 (Table 2) for both ZST and supervised directions for each dataset**. We report the average accuracy of three seeds and all the supervised or zero-shot directions. "Pre-Src" and "Pre-Tgt" indicate the layer-wise source and target language recognition for a PreNorm system (#3, #5, or #7), while "Post-Src" and "Post-Tgt" denote similary for a PostNorm system (#4, #6, or #8). "L1" to "L6" are 6 encoder layers and "L7" to "L12" are 6 decoder layers.

Layer Norm	Direction	S-ENC-T-DEC w/ Res.				T-ENC w/ Res.				S-ENC-T-DEC w/o Res.				T-ENC w/o Res.			
		1	10	20	avg.	1	10	20	avg.	1	10	20	avg.	1	10	20	avg.
Pre.	ar-de	5.3	5.9	5.2	5.5	10.0	11.0	10.0	10.3	9.8	8.4	9.5	9.2	11.4	8.3	10.8	10.2
	ar-fr	17.5	16.1	17.2	16.9	16.3	19.9	18.3	18.2	19.9	20.3	20.8	20.3	20.9	18.6	20.4	20.0
	ar-nl	8.6	6.3	7.9	7.6	13.2	13.1	12.6	13.0	13.3	14.0	12.2	13.2	13.5	12.8	13.6	13.3
	ar-ru	8.1	9.1	9.5	8.9	14.8	16.2	15.9	15.6	13.0	10.9	13.0	12.3	19.6	17.8	19.6	19.0
	ar-zh	12.7	13.4	13.8	13.3	28.1	28.1	27.3	27.8	25.1	19.8	24.4	23.1	31.2	31.0	31.3	31.2
	de-ar	3.6	3.3	2.5	3.1	5.6	5.0	3.9	4.8	6.9	6.4	5.6	6.3	6.4	5.1	3.3	4.9
	de-fr	15.5	16.0	16.2	15.9	5.1	5.7	3.8	4.9	18.8	17.6	18.9	18.4	5.9	4.7	5.2	5.3
	de-nl	19.4	15.9	18.8	18.0	12.4	8.9	8.6	10.0	21.4	20.4	20.7	20.8	9.1	7.1	7.7	8.0
	de-ru	6.0	6.1	5.8	6.0	5.0	5.6	3.7	4.8	9.4	9.2	9.0	9.2	6.4	4.5	3.8	4.9
	de-zh	7.6	9.5	8.8	8.6	15.6	12.4	11.9	13.3	14.4	12.8	13.2	13.5	16.4	4.1	6.0	8.8
	fr-ar	9.5	7.5	8.5	8.5	15.5	16.2	13.2	15.0	15.4	13.1	14.4	14.3	18.5	16.5	15.8	16.9
	fr-de	10.4	10.6	11.6	10.9	6.3	7.2	4.9	6.1	14.1	11.0	15.2	13.4	4.5	4.6	4.0	4.4
	fr-nl	17.5	13.7	18.0	16.4	16.0	12.5	13.2	13.9	20.5	19.9	20.2	20.2	11.1	9.1	8.6	9.6
	fr-ru	8.8	8.4	9.3	8.8	12.1	12.8	10.9	11.9	13.3	9.2	12.0	11.5	16.5	7.4	8.4	10.8
	fr-zh	14.3	13.1	15.3	14.2	31.2	30.0	28.0	29.7	27.9	21.0	25.8	24.9	34.1	16.0	27.9	26.0
	nl-ar	2.6	2.0	1.7	2.1	5.2	5.6	5.3	5.4	4.3	5.8	4.2	4.8	5.5	5.0	5.0	5.2
	nl-de	14.3	14.4	13.9	14.2	12.8	13.9	11.3	12.7	16.9	14.8	18.3	16.7	13.8	6.9	10.9	10.5
	nl-fr	18.3	17.4	18.5	18.1	13.1	16.1	12.4	13.9	21.5	19.9	22.3	21.2	15.0	7.1	13.8	12.0
	nl-ru	4.2	4.4	3.4	4.0	9.5	9.8	8.6	9.3	7.2	6.5	7.3	7.0	10.3	6.6	7.3	8.1
	nl-zh	2.2	3.2	3.0	2.8	10.8	10.0	10.4	10.4	7.0	8.0	6.3	7.1	11.1	7.5	10.0	9.5
	ru-ar	9.7	7.6	7.6	8.3	15.6	16.1	14.6	15.4	15.9	13.3	14.0	14.4	18.6	19.1	18.0	18.6
	ru-de	7.7	9.1	7.2	8.0	8.5	10.0	6.0	8.2	10.5	10.0	10.9	10.5	8.4	5.6	6.8	6.9
	ru-fr	18.1	17.5	17.4	17.7	18.1	20.5	17.6	18.7	19.9	19.5	20.7	20.0	22.4	17.4	21.1	20.3
	ru-nl	10.2	8.6	9.9	9.6	11.5	11.7	9.5	10.9	13.0	13.1	12.4	12.8	12.7	8.2	10.1	10.3
	ru-zh	11.3	11.6	12.5	11.8	28.4	28.3	27.6	28.1	25.3	17.7	21.6	21.5	31.9	20.0	30.7	27.5
	zh-ar	9.1	7.6	7.2	8.0	15.2	16.6	14.5	15.4	15.6	12.7	15.1	14.5	18.4	18.8	18.7	18.6
	zh-fr	16.7	15.6	16.4	16.2	20.1	21.4	18.4	20.0	20.9	19.3	20.6	20.3	23.5	23.3	23.7	23.5
	zh-de	4.7	5.8	5.4	5.3	7.8	8.1	7.0	7.6	7.5	6.9	7.1	7.2	8.6	8.6	8.8	8.7
	zh-nl	6.9	5.4	6.0	6.1	8.6	8.6	8.2	8.5	8.5	8.0	8.0	8.2	9.1	9.2	8.8	9.0
	zh-ru	6.9	8.2	7.8	7.6	13.7	15.7	12.9	14.1	12.8	10.0	11.8	11.5	18.7	19.8	19.7	19.4
	avg.	10.3	9.8	10.2	**10.1**	13.5	13.9	12.4	**13.3**	15.0	13.3	14.5	**14.3**	15.1	11.7	13.3	**13.4**
Post.	ar-de	11.4	11.0	10.3	10.9	10.1	10.4	9.9	10.1	10.1	11.9	9.9	10.6	11.0	11.0	10.0	10.7
	ar-fr	20.7	23.2	20.3	21.4	16.2	18.7	19.3	18.1	20.7	24.0	19.2	21.3	20.4	21.8	15.9	19.4
	ar-nl	13.3	13.7	12.5	13.2	12.8	13.5	13.3	13.2	13.4	14.4	12.5	13.4	13.2	13.9	13.0	13.4
	ar-ru	16.9	18.7	16.1	17.2	17.4	17.2	18.6	17.7	13.5	19.1	14.7	15.8	20.4	20.7	18.7	19.9
	ar-zh	28.6	29.4	29.2	29.1	29.2	30.4	30.3	30.0	26.1	30.7	27.4	28.1	32.9	32.9	31.9	32.6
	de-ar	7.2	7.2	6.6	7.0	5.7	5.6	5.8	5.7	6.9	7.6	7.6	7.4	4.4	4.1	3.1	3.9
	de-fr	17.6	19.3	18.2	18.4	5.1	6.6	5.8	5.8	17.3	20.3	17.3	18.3	5.4	7.9	4.1	5.8
	de-nl	21.4	21.8	20.4	21.2	9.1	9.5	7.9	8.8	20.0	22.3	20.5	20.9	9.7	11.9	7.1	9.6
	de-ru	12.3	13.8	12.8	13.0	6.0	6.3	7.2	6.5	10.1	13.3	10.5	11.3	5.2	4.0	3.7	4.3
	de-zh	16.1	16.9	16.5	16.5	8.9	15.3	15.0	13.1	11.2	16.9	13.5	13.9	14.1	11.1	3.1	9.4
	fr-ar	17.9	17.8	18.9	18.2	16.4	17.1	16.4	16.6	14.6	19.5	16.3	16.8	16.4	16.6	14.8	15.9
	fr-de	15.0	17.3	17.0	16.4	5.4	6.7	6.5	6.2	13.1	17.0	13.5	14.5	4.9	7.0	4.8	5.6
	fr-nl	21.4	21.8	20.3	21.2	11.3	13.3	11.6	12.1	20.6	22.7	20.5	21.3	11.6	14.1	10.1	11.9
	fr-ru	17.7	19.5	15.9	17.7	16.7	13.3	18.5	16.2	12.9	20.7	13.3	15.6	10.9	15.5	13.3	13.2
	fr-zh	30.5	32.0	31.8	31.4	29.8	32.0	31.4	31.1	25.9	32.5	28.4	28.9	31.7	32.0	30.3	31.3
	nl-ar	5.3	5.9	5.6	5.6	6.0	5.3	5.8	5.7	5.2	6.1	6.4	5.9	5.0	5.2	4.5	4.9
	nl-de	17.9	19.7	19.1	18.9	10.9	12.8	10.5	11.4	16.5	19.8	17.1	17.8	9.0	10.4	10.4	9.9
	nl-fr	21.1	22.5	21.2	21.6	13.8	13.4	13.0	13.4	21.2	22.9	19.6	21.2	10.1	12.6	9.5	10.7
	nl-ru	10.0	11.2	10.2	10.5	9.7	9.1	8.8	9.2	8.4	10.9	8.6	9.3	8.6	7.6	8.2	8.1
	nl-zh	9.6	11.1	9.6	10.1	10.2	10.4	10.0	10.2	5.4	11.1	7.3	7.9	9.9	9.9	7.5	9.1
	ru-ar	18.7	18.7	18.2	18.5	16.9	17.9	17.5	17.4	14.8	19.7	16.2	16.9	17.9	18.9	17.0	17.9
	ru-de	12.9	12.9	12.9	12.9	8.7	8.1	9.0	8.6	10.8	13.3	10.5	11.5	8.6	9.2	7.9	8.6
	ru-fr	21.5	24.0	21.2	22.2	19.4	17.9	19.0	18.8	20.1	24.8	19.0	21.3	16.8	22.0	13.8	17.5
	ru-nl	13.0	13.6	12.7	13.1	10.9	11.8	12.4	11.7	13.3	14.2	13.0	13.5	11.0	12.0	9.7	10.9
	ru-zh	27.6	29.8	28.6	28.7	30.1	30.4	30.6	30.4	23.6	30.2	24.6	26.1	32.5	32.2	29.0	31.2
	zh-ar	18.0	17.4	17.3	17.6	16.9	17.5	17.1	17.2	16.3	19.3	17.0	17.5	19.1	19.8	19.4	19.4
	zh-fr	20.2	21.3	20.2	20.6	21.4	22.3	21.5	21.7	20.5	24.1	18.3	21.0	23.1	24.4	24.5	24.0
	zh-de	8.6	9.1	8.8	8.8	7.3	7.4	7.1	7.3	8.3	9.9	7.5	8.6	8.7	8.5	8.0	8.4
	zh-nl	8.7	8.5	8.1	8.4	8.9	8.7	8.4	8.7	8.9	9.0	8.1	8.7	8.9	9.3	9.0	9.1
	zh-ru	15.3	15.8	14.1	15.1	16.7	17.8	17.6	17.2	13.3	17.8	12.8	14.6	20.2	20.5	20.2	20.3
	avg.	16.5	17.5	16.5	**16.8**	13.6	14.2	14.2	**14.0**	14.8	18.2	15.0	**16.0**	14.1	14.9	12.8	**13.9**

Table 5: **BLEU scores of OPUS in ZST directions**. Scores in **bold** are the results reported in Table 2. "1," "10," and "20" indicates three random seeds. "Res." indicates the residual connection of self-attention in the 4th encoder layer.

Layer Norm	Direction	S-ENC-T-DEC w/ Res.				T-ENC w/ Res.				S-ENC-T-DEC w/o Res.				T-ENC w/o Res.			
		1	10	20	avg.	1	10	20	avg.	1	10	20	avg.	1	10	20	avg.
Pre.	en-ar	23.6	24.1	23.2	23.6	23.7	23.9	24.1	23.9	24.0	23.2	23.1	23.4	22.8	23.8	23.8	23.5
	ar-en	37.6	37.1	37.3	37.3	37.5	37.1	37.5	37.4	37.4	37.2	36.9	37.2	36.4	36.7	37.0	36.7
	en-de	29.7	30.1	30.4	30.1	30.4	29.6	30.4	30.1	30.1	30.1	30.1	30.1	30.3	30.5	30.7	30.5
	de-en	34.3	34.5	34.2	34.3	34.5	34.1	34.3	34.3	35.0	34.7	34.3	34.7	33.8	34.1	34.4	34.1
	en-fr	33.5	33.7	33.6	33.6	33.4	33.8	33.6	33.6	33.7	33.1	33.8	33.5	33.0	33.6	33.1	33.2
	fr-en	35.6	35.4	35.3	35.4	35.0	35.0	35.5	35.2	35.6	35.2	35.1	35.3	34.4	35.2	35.0	34.9
	en-nl	27.7	28.4	28.2	28.1	28.4	27.9	28.3	28.2	27.6	28.0	27.9	27.8	28.1	28.1	28.0	28.1
	nl-en	31.3	30.8	31.2	31.1	30.9	30.7	30.8	30.8	31.0	30.8	31.0	30.9	30.4	30.9	30.5	30.6
	en-ru	29.2	29.7	29.6	29.5	29.4	29.8	29.8	29.7	29.5	29.1	29.6	29.4	29.4	29.9	29.2	29.5
	ru-en	35.2	34.6	35.0	34.9	34.7	34.6	35.0	34.8	35.2	34.8	35.1	35.0	34.3	34.8	34.7	34.6
	en-zh	40.7	40.8	40.9	40.8	40.6	40.3	40.7	40.5	40.7	40.4	40.6	40.6	39.6	40.7	40.6	40.3
	zh-en	46.2	46.1	45.9	46.1	46.1	46.1	46.2	46.1	46.2	45.9	45.8	46.0	45.6	46.4	46.3	46.1
	avg.	33.7	33.8	33.7	**33.7**	33.7	33.6	33.9	**33.7**	33.8	33.5	33.6	**33.7**	33.2	33.7	33.6	**33.5**
Post.	en-ar	23.9	23.4	23.7	23.7	24.6	24.4	24.3	24.4	23.7	23.8	23.8	23.8	24.0	23.8	24.0	23.9
	ar-en	37.8	37.3	37.5	37.5	37.8	37.5	37.2	37.5	37.7	37.2	37.6	37.5	37.8	37.3	37.7	37.6
	en-de	30.8	31.0	29.3	30.4	31.2	29.9	31.2	30.8	31.1	30.5	31.2	30.9	31.1	30.5	31.5	31.0
	de-en	34.6	34.6	34.8	34.7	34.9	34.6	34.7	34.7	34.8	34.6	34.7	34.7	34.4	34.6	34.4	34.5
	en-fr	33.9	33.4	34.1	33.8	34.1	33.8	33.9	33.9	33.5	33.5	33.2	33.4	33.7	33.8	33.6	33.7
	fr-en	35.5	35.6	35.4	35.5	35.6	35.7	35.4	35.6	35.0	35.5	35.2	35.2	35.3	35.3	35.5	35.4
	en-nl	27.8	28.4	28.2	28.1	27.9	28.8	28.3	28.3	28.0	27.9	28.3	28.1	27.7	27.9	28.4	28.0
	nl-en	31.5	30.9	31.2	31.2	31.3	30.9	31.4	31.2	30.8	30.8	30.7	30.8	31.1	31.1	30.9	31.0
	en-ru	29.4	29.6	29.9	29.6	30.1	29.8	30.0	30.0	29.9	30.0	29.2	29.7	30.0	29.5	29.5	29.7
	ru-en	35.1	34.6	35.1	34.9	34.9	34.9	35.2	35.0	34.8	34.9	35.2	35.0	34.8	34.8	35.0	34.9
	en-zh	41.2	40.9	40.9	41.0	41.2	40.9	40.8	41.0	40.8	40.5	40.7	40.7	40.7	40.7	41.0	40.8
	zh-en	46.4	46.0	46.1	46.2	46.7	46.3	46.2	46.4	46.1	46.3	46.1	46.2	46.7	46.6	46.0	46.4
	avg.	34.0	33.8	33.9	**33.9**	34.2	34.0	34.1	**34.1**	33.9	33.8	33.8	**33.8**	33.9	33.8	34.0	**33.9**

Table 6: **BLEU scores of OPUS in supervised directions**. Scores in **bold** are the results reported in Table 2. "1," "10," and "20" indicates three random seeds. "Res." indicates the residual connection of self-attention in the 4th encoder layer.

Layer Norm	Direction	S-ENC-T-DEC w/ Res.				T-ENC w/ Res.				S-ENC-T-DEC w/o Res.				T-ENC w/o Res.			
		1	10	20	avg.	1	10	20	avg.	1	10	20	avg.	1	10	20	avg.
Pre.	it-nl	5.2	3.7	4.3	4.4	13.4	14.4	14.0	13.9	6.4	3.6	13.8	7.9	16.3	17.7	17.2	17.1
	nl-it	5.5	4.3	4.3	4.7	13.9	14.7	14.4	14.3	6.1	4.6	10.8	7.2	15.5	17.0	17.1	16.5
	it-ro	5.5	5.7	5.1	5.4	13.4	13.5	14.4	13.8	7.8	7.4	14.2	9.8	16.0	16.6	16.9	16.5
	ro-it	7.2	5.5	5.3	6.0	14.9	15.1	15.4	15.1	7.1	4.3	11.4	7.6	17.8	18.1	18.4	18.1
	nl-ro	4.5	4.9	4.2	4.5	12.1	12.5	12.4	12.3	6.1	7.1	11.8	8.3	12.8	14.1	14.1	13.7
	ro-nl	4.4	4.3	3.9	4.2	12.1	13.4	12.5	12.7	5.6	3.1	12.4	7.0	15.1	16.1	15.6	15.6
	avg.	5.4	4.7	4.5	**4.9**	13.3	13.9	13.9	**13.7**	6.5	5.0	12.4	**8.0**	15.6	16.6	16.6	**16.2**
Post.	it-nl	13.7	11.8	13.1	12.9	15.9	16.3	17.0	16.4	17.7	18.3	17.4	17.8	18.4	18.0	18.6	18.3
	nl-it	14.5	12.8	12.2	13.2	15.7	17.0	16.1	16.3	18.0	18.5	18.4	18.3	17.9	18.3	18.3	18.2
	it-ro	12.3	11.2	12.4	12.0	14.8	14.3	15.8	15.0	17.0	17.3	17.0	17.1	17.9	17.8	18.2	18.0
	ro-it	14.6	13.7	13.0	13.8	17.2	16.8	17.5	17.2	19.5	20.0	20.0	19.8	19.2	19.8	20.8	19.9
	nl-ro	11.1	10.4	10.2	10.6	13.5	13.4	13.6	13.5	14.9	14.9	14.7	14.8	15.4	15.2	15.5	15.4
	ro-nl	12.3	10.9	12.2	11.8	14.5	15.0	15.2	14.9	16.5	16.6	16.0	16.4	16.9	16.2	17.1	16.7
	avg.	13.1	11.8	12.2	**12.4**	15.3	15.5	15.9	**15.5**	17.3	17.6	17.3	**17.4**	17.6	17.6	18.1	**17.8**

Table 7: **BLEU scores of IWSLT in ZST directions**. Scores in **bold** are the results reported in Table 2. "1," "10," and "20" indicates three random seeds. "Res." indicates the residual connection of self-attention in the 4th encoder layer.

Layer Norm	Direction	S-ENC-T-DEC w/ Res.				T-ENC w/ Res.				S-ENC-T-DEC w/o Res.				T-ENC w/o Res.			
		1	10	20	avg.	1	10	20	avg.	1	10	20	avg.	1	10	20	avg.
Pre.	en-it	33.9	33.8	33.6	33.8	33.7	33.4	33.7	33.6	33.6	32.9	33.3	33.3	32.4	33.3	33.4	33.0
	it-en	37.5	37.1	37.1	37.2	37.4	37.2	37.0	37.2	35.8	36.3	36.5	36.2	35.8	36.7	36.5	36.3
	en-nl	29.6	29.5	29.4	29.5	29.6	29.5	29.6	29.6	29.2	29.7	29.5	29.5	29.0	29.2	29.2	29.1
	nl-en	31.9	32.4	32.0	32.1	32.0	32.1	31.9	32.0	30.9	31.3	31.7	31.3	31.2	31.5	31.5	31.4
	en-ro	24.4	25.1	25.1	24.9	25.2	25.1	25.4	25.2	24.4	24.6	24.4	24.5	24.6	24.7	24.6	24.6
	ro-en	31.3	31.6	31.3	31.4	32.1	31.6	31.4	31.7	30.3	30.7	30.9	30.6	30.3	31.2	31.2	30.9
	avg.	31.4	31.6	31.4	**31.5**	31.7	31.5	31.5	**31.6**	30.7	30.9	31.1	**30.9**	30.6	31.1	31.1	**30.9**
Post.	en-it	33.9	33.3	33.5	33.6	33.8	34.0	33.5	33.8	33.1	33.2	32.6	33.0	32.4	32.6	33.4	32.8
	it-en	37.1	36.9	37.0	37.0	37.1	37.1	36.9	37.0	35.7	35.4	36.1	35.7	36.4	35.7	35.8	36.0
	en-nl	29.6	30.1	30.1	29.9	30.4	30.4	30.0	30.3	29.2	29.0	29.0	29.1	29.2	29.0	29.5	29.2
	nl-en	31.9	32.0	31.6	31.8	31.3	31.9	31.8	31.7	31.0	31.1	31.7	31.3	30.9	30.7	31.3	31.0
	en-ro	25.4	25.2	24.6	25.1	25.3	25.2	25.5	25.3	24.7	25.0	24.6	24.8	24.4	24.4	25.0	24.6
	ro-en	31.5	31.6	31.6	31.6	30.8	31.4	31.1	31.1	30.4	29.6	30.8	30.3	30.4	30.1	30.4	30.3
	avg.	31.6	31.5	31.4	**31.5**	31.5	31.7	31.5	**31.5**	30.7	30.6	30.8	**30.7**	30.6	30.4	30.9	**30.6**

Table 8: **BLEU scores of IWSLT in supervised directions**. Scores in **bold** are the results reported in Table 2. "1," "10," and "20" indicates three random seeds. "Res." indicates the residual connection of self-attention in the 4th encoder layer.

Layer Norm	Direction	S-ENC-T-DEC w/ Res.				T-ENC w/ Res.				S-ENC-T-DEC w/o Res.				T-ENC w/o Res.			
		1	10	20	avg.	1	10	20	avg.	1	10	20	avg.	1	10	20	avg.
Pre.	es-de	23.2	22.0	16.1	20.4	26.7	26.9	27.3	27.0	6.2	14.1	11.2	10.5	24.9	28.5	28.3	27.2
	de-es	30.3	30.0	27.6	29.3	32.4	32.0	32.3	32.2	15.5	25.7	18.7	20.0	32.9	33.1	33.4	33.1
	es-fr	35.0	35.6	34.0	34.9	38.8	38.8	39.3	39.0	27.8	29.8	28.2	28.6	39.9	39.8	39.9	39.9
	fr-es	36.0	35.5	32.8	34.8	38.6	38.7	38.7	38.7	18.7	30.7	22.3	23.9	39.7	39.7	40.0	39.8
	es-nl	22.7	23.0	14.2	20.0	26.4	26.3	26.3	26.3	7.0	12.8	15.0	11.6	23.2	27.7	27.5	26.1
	nl-es	27.2	27.1	24.9	26.4	29.1	29.1	29.1	29.1	13.9	23.0	16.9	17.9	29.6	29.7	29.8	29.7
	de-fr	28.6	28.1	26.9	27.9	31.4	31.3	31.7	31.5	21.9	23.0	22.5	22.5	31.9	32.3	32.2	32.1
	fr-de	23.5	22.0	15.9	20.5	26.3	26.5	26.8	26.5	6.3	14.3	11.5	10.7	25.0	28.1	28.2	27.1
	de-nl	23.2	23.4	15.0	20.5	26.3	26.2	26.0	26.2	7.0	12.8	16.2	12.0	22.5	27.5	27.2	25.7
	nl-de	21.4	20.3	14.3	18.7	23.2	23.8	23.5	23.5	6.4	13.3	11.9	10.5	21.6	24.6	24.6	23.6
	fr-nl	22.9	23.3	14.1	20.1	26.0	25.9	25.8	25.9	6.8	12.2	15.3	11.4	21.6	27.4	27.1	25.4
	nl-fr	26.0	25.9	25.0	25.6	28.1	28.3	28.2	28.2	19.9	20.9	19.9	20.2	28.9	28.8	28.7	28.8
	avg.	26.7	26.4	21.7	**24.9**	29.4	29.5	29.6	**29.5**	13.1	19.4	17.5	**16.7**	28.5	30.6	30.6	**29.9**
Post.	es-de	26.0	26.9	26.8	26.6	28.2	28.4	28.7	28.4	26.1	26.3	26.1	26.2	28.7	28.7	28.7	28.7
	de-es	32.3	32.6	32.1	32.3	33.2	33.7	33.5	33.5	32.7	31.9	32.1	32.2	33.5	33.3	33.5	33.4
	es-fr	37.7	38.8	37.5	38.0	40.2	40.0	40.1	40.1	37.9	37.8	37.7	37.8	40.1	39.9	40.5	40.2
	fr-es	37.8	38.5	38.2	38.2	40.0	39.9	40.1	40.0	38.4	37.7	38.0	38.0	39.7	39.7	40.1	39.8
	es-nl	25.6	26.0	26.2	25.9	27.9	27.7	27.8	27.8	26.0	25.7	25.5	25.7	27.8	28.0	27.9	27.9
	nl-es	29.3	29.3	29.1	29.2	29.8	30.0	29.6	29.8	29.4	29.0	29.2	29.2	29.7	29.8	29.8	29.8
	de-fr	30.6	31.7	30.8	31.0	32.8	32.8	33.1	32.9	31.0	30.7	30.8	30.8	32.9	32.4	33.3	32.9
	fr-de	25.9	26.4	26.6	26.3	27.8	28.6	28.8	28.4	26.3	26.0	25.1	25.8	28.2	28.5	28.3	28.3
	de-nl	25.8	26.0	25.9	25.9	27.5	27.7	27.5	27.6	25.7	25.6	25.5	25.6	27.8	27.6	27.5	27.6
	nl-de	23.5	23.4	23.9	23.6	24.2	24.6	24.4	24.4	23.6	23.5	23.2	23.4	24.4	24.5	24.5	24.5
	fr-nl	25.3	25.8	25.6	25.6	27.4	27.4	27.3	27.4	25.5	25.5	25.3	25.4	27.8	27.6	27.5	27.6
	nl-fr	28.1	28.4	27.9	28.1	29.3	29.0	29.3	29.2	28.3	28.0	27.9	28.1	29.2	29.1	29.3	29.2
	avg.	29.0	29.5	29.2	**29.2**	30.7	30.8	30.9	**30.8**	29.2	29.0	28.9	**29.0**	30.8	30.8	30.9	**30.8**

Table 9: **BLEU scores of Europarl in ZST directions**. Scores in **bold** are the results reported in Table 2. "1," "10," and "20" indicates three random seeds. "Res." indicates the residual connection of self-attention in the 4th encoder layer.

Layer Norm	Direction	S-ENC-T-DEC w/ Res.				T-ENC w/ Res.				S-ENC-T-DEC w/o Res.				T-ENC w/o Res.			
		1	10	20	avg.	1	10	20	avg.	1	10	20	avg.	1	10	20	avg.
Pre.	en-de	28.0	28.0	28.3	28.1	28.2	28.2	28.4	28.3	28.0	28.1	28.4	28.2	28.5	28.5	28.3	28.4
	de-en	35.2	35.1	35.3	35.2	35.1	35.0	35.1	35.1	34.9	35.0	35.0	35.0	34.8	35.1	35.0	35.0
	en-es	37.6	37.4	37.4	37.5	37.5	37.4	37.7	37.5	37.5	37.5	37.4	37.5	37.5	37.5	37.3	37.4
	es-en	39.3	38.9	39.0	39.1	39.0	39.0	38.9	39.0	38.8	39.0	39.1	39.0	38.6	39.0	38.9	38.8
	en-fr	36.2	36.6	36.5	36.4	36.5	36.4	36.8	36.6	36.3	36.4	36.5	36.4	36.7	36.7	36.2	36.5
	fr-en	38.2	38.2	38.0	38.1	38.0	38.2	38.0	38.1	38.0	37.9	38.2	38.0	37.8	38.2	38.0	38.0
	en-nl	28.5	28.8	28.7	28.7	28.8	28.7	28.6	28.7	28.5	28.6	28.6	28.6	28.3	28.6	28.3	28.4
	nl-en	31.7	31.6	31.5	31.6	31.5	31.7	31.9	31.7	31.6	31.3	31.6	31.5	31.3	31.7	31.6	31.5
	avg.	34.3	34.3	34.3	**34.3**	34.3	34.3	34.4	**34.4**	34.2	34.2	34.4	**34.3**	34.2	34.4	34.2	**34.3**
Post.	en-de	28.4	28.4	28.7	28.5	28.6	28.7	29.0	28.8	28.5	28.2	28.4	28.4	28.7	28.5	28.3	28.5
	de-en	35.2	35.0	35.5	35.2	34.8	35.1	34.9	34.9	35.2	35.2	35.0	35.1	35.1	35.1	34.7	35.0
	en-es	37.6	37.8	37.5	37.6	37.6	37.7	37.6	37.6	37.6	37.5	37.6	37.6	37.3	37.4	37.5	37.4
	es-en	39.4	39.0	39.0	39.1	39.0	39.3	38.8	39.0	39.2	38.9	39.1	39.1	39.0	39.1	39.1	39.1
	en-fr	36.8	36.8	36.4	36.7	36.8	36.7	37.0	36.8	36.6	36.5	37.1	36.7	36.9	36.8	36.7	36.8
	fr-en	38.3	38.2	38.4	38.3	38.2	38.2	38.4	38.3	38.2	38.1	38.2	38.2	38.1	38.3	37.9	38.1
	en-nl	28.8	28.8	28.6	28.7	28.7	28.7	28.9	28.8	28.6	28.6	28.9	28.7	28.7	28.7	28.5	28.6
	nl-en	31.5	31.6	31.7	31.6	32.1	31.7	31.7	31.8	31.7	31.9	31.5	31.7	31.7	31.4	31.4	31.5
	avg.	34.5	34.5	34.5	**34.5**	34.5	34.5	34.5	**34.5**	34.5	34.4	34.5	**34.4**	34.4	34.4	34.3	**34.4**

Table 10: **BLEU scores of Europarl supervised directions**. Scores in **bold** are the results reported in Table 2. "1," "10," and "20" indicates three random seeds. "Res." indicates the residual connection of self-attention in the 4th encoder layer.

#	Layer Norm	Language Tag	Res.	Zero-shot			Supervised		
				OPUS	IWSLT	Europarl	OPUS	IWSLT	Europarl
0		*Pivot*		55.8	64.6	73.8	-	-	-
1	**PreNorm**	S-ENC-T-DEC	w/	35.9	34.6	66.5	63.8	70.6	74.9
2	**PostNorm**	S-ENC-T-DEC	w/	**49.1**	**51.2**	**73.0**	64.1	70.6	75.0
3	**PreNorm**	T-ENC	w/	42.5	53.0	73.0	63.7	70.6	74.9
4	**PostNorm**	T-ENC	w/	**43.8**	**56.0**	**73.8**	64.0	70.7	75.0
5	**PreNorm**	S-ENC-T-DEC	w/o	44.5	41.7	50.3	63.7	70.0	74.8
6	**PostNorm**	S-ENC-T-DEC	w/o	**47.6**	**60.8**	**72.9**	64.0	69.7	74.9
7	**PreNorm**	T-ENC	w/o	42.5	57.1	72.5	63.6	69.9	74.8
8	**PostNorm**	T-ENC	w/o	**43.1**	**60.2**	**73.8**	64.0	69.7	74.9

Table 11: **BLEURT scores**. We report the mean of three seeds and all the translation directions. "Res." indicates the residual connection of self-attention in the 4th encoder layer. We mark better scores between PreNorm and PostNorm in **bold** for ZST.

A For every submission:

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

☑ A2. Did you discuss any potential risks of your work?
Section "Ethical Considerations"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
Only Grammarly was utilized for grammar correction, and there is no originally machine-generated text in the paper.

B ☑ Did you use or create scientific artifacts?

Section 2.1 and 2.2 Appendix A and B

☑ B1. Did you cite the creators of artifacts you used?
Section 2.1 and 2.2 Appendix A and B

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

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

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Following previous work of neural machine translation, we directly used OPUS, IWSLT, and Europarl datasets without further checking filtering in order to conduct fair comparisons.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2.1 and Appendix A and B

C ☑ Did you run computational experiments?

Section 2.1, 2.2, 2.3, and 2.4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 2.1 and Appendix A

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 2.1 Appendix A and B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 2.2, 2.3, and 2.4 Appendix C, E, F, and G

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2.1 Appendix A and B

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

Left blank.

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

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

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

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

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