The Effect of Language Diversity When Fine-Tuning Large Language Models for Translation

Anonymous ACL submission

Abstract

Prior research diverges on language diversity in LLM fine-tuning: Some studies report benefits while others find no advantages. Through controlled fine-tuning experiments across 132 translation directions, we systematically resolve these disparities. We find that expanding language diversity during fine-tuning improves translation quality for both unsupervised and—surprisingly—supervised pairs, despite less diverse models being fine-tuned exclusively on these supervised pairs. However, benefits plateau or decrease beyond a certain diversity threshold. We show that increased language diversity creates more language-agnostic representations. These representational adaptations help explain the improved performance in models fine-tuned with greater diversity.[1]

1 Introduction

General-purpose LLMs like LLAMA 3 (Grattafiori et al., 2024) show promise for machine translation but require targeted fine-tuning beyond their incidental bilingualism (Briakou et al., 2023) to match the performance of specialized translation systems. Through fine-tuning approaches ranging from two-stage methods (Li et al., 2024; Zeng et al., 2024; Stap et al., 2024) to more sophisticated optimization techniques (Xu et al., 2025; Zhu et al., 2024b), LLMs such as TOWER (Alves et al., 2024) now outperform traditional NMT systems (Kocmi et al., 2024; Deutsch et al., 2025).

Current research presents conflicting evidence on multilingual fine-tuning strategies. Some studies show that scaling the number of tasks or languages during instruction tuning improves (cross-lingual) generalization (Wang et al., 2022; Muennighoff et al., 2023; Dang et al., 2024), while others report that just 1–3 fine-tuning languages effectively trigger cross-lingual transfer (Kew et al., 2024; Zhu et al., 2024a). Recent *inference-only* experiments by Richburg and Carpuat (2024) across 132 translation directions highlight this uncertainty, showing variance in translation quality with off-target generations for non-English sources and inconsistent performance across languages. While non-English over-tokenization and typological distance provide partial explanations, controlled fine-tuning experiments on the effects of language diversity *during fine-tuning* remain unexplored.

We address these conflicting findings through systematic experimentation with varying translation directions, measuring effects on both seen and unseen language pairs. Through controlled fine-tuning across 132 translation directions, we demonstrate that increasing language diversity consistently improves translation quality in all categories. Counterintuitively, models fine-tuned on the most diverse language sets outperform others *even on fully supervised language pairs* that less diverse models are specifically optimized to handle. However, experiments with even larger language sets (272 directions) reveal that benefits plateau or decrease beyond a certain diversity threshold. Analysis of model activations shows that fine-tuning on diverse language directions creates more target language-agnostic representations in middle layers, explaining the performance improvements in our most diverse models.

2 Fine-tuning and evaluation design

Following Richburg and Carpuat (2024), we categorize our language pairs into three groups based on their presence in the fine-tuning data of the TOWER model we build upon: *fully supervised* (pairs between `de`, `en`, `ko`, `nl`, `ru` and `zh`), *zero-shot* (pairs involving `cs`, `is`, `ja`, `pl`, `sv` and `uk`), and *partially supervised* (pairs combining supervised and zero-shot languages). This yields 132 translation directions across 12 typologically diverse languages with varying pre-training represen-

[1]We will release our code and models upon acceptance.

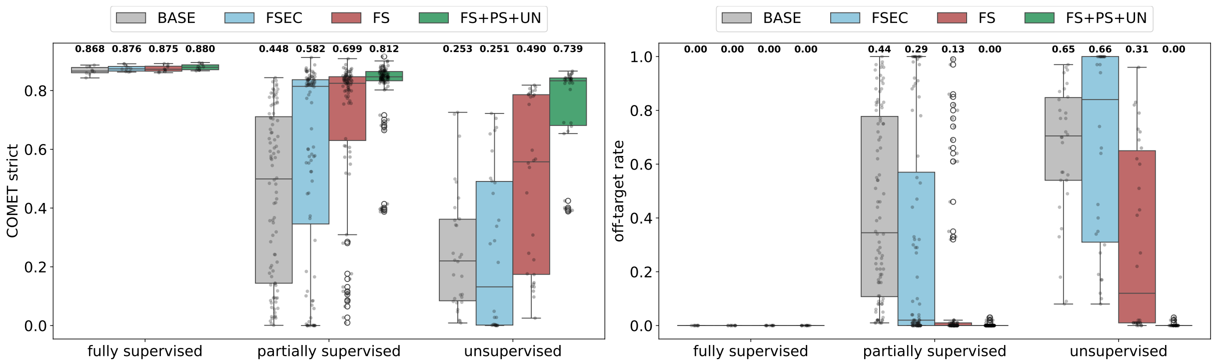

Figure 1: COMET-STRICT scores (left) and off-target rates (right) for BASE (no fine-tuning), FSEC (English-centric), FS (seen directions), FS+PS+UN (all directions), evaluated on *fully supervised* (de/en/ko/nl/ru/zh pairs), *unsupervised* (cs/is/ja/pl/sv/uk pairs), and *partially supervised* (combining supervised and unsupervised) language pairs. Numbers above bars show mean scores. Training on more diverse sets improves *all* categories, with FS+PS+UN achieving best COMET-STRICT scores even for fully supervised pairs. FS substantially reduces off-target rates for unsupervised directions compared to BASE and FSEC, despite these pairs being *absent* from its fine-tuning data.

tation, enabling comprehensive assessment across different data conditions (see Appendix A).

Fine-tuning setups We compare the following incremental fine-tuning approaches using the TOWER family of models, which are built on LLAMA 2 and underwent continued pre-training with a mixture of monolingual and parallel data:
BASE: TOWERBASE-7B model without task-specific fine-tuning, serving as our baseline.
FSEC: BASE fine-tuned only on fully supervised English-centric translation directions (10 directions), representing minimal supervision.
FS: BASE fine-tuned on all fully supervised language directions (30 directions), extending beyond English-centric pairs to investigate transfer learning between diverse language combinations.
FS+PS+UN: BASE fine-tuned on fully supervised, partially supervised, and unsupervised directions (132 directions), maximizing language diversity to investigate cross-lingual transfer effects.

This controlled experimental design allows us to systematically evaluate how increasing language diversity during fine-tuning affects both supervised and unsupervised translation directions, moving beyond aggregate scores to understand performance patterns across specific language groups.

Data We fine-tune on NTREX-128 (Federmann et al., 2022), a high-quality dataset of 1,997 multi-parallel professionally translated sentences designed for machine translation evaluation.[2] For evaluation, we use the FLORES-200 (Team et al., 2022) devtest set, which provides multi-parallel data for controlled cross-language comparison.

Metrics Our primary metric is COMET-STRICT, a modified version of COMET (Rei et al., 2020) that assigns zero scores to off-target translations, following recommendations by Zouhar et al. (2024).[3] We also report off-target rates, measured using FASTTEXT (Joulin et al., 2017, 2016) language identification.[4] Optimization and inference details are provided in Appendix B.

3 Results

Increased diversity leads to better performance Figure 1 (left) demonstrates that expanding language diversity during fine-tuning yields consistent performance improvements across all language pair categories. The COMET-STRICT scores show a clear progression from BASE to FSEC to FS to FS+PS+UN models, with the most diverse model achieving the highest scores in every category. Surprisingly, the FS+PS+UN model (fine-tuned on *all* 132 directions) outperforms specialized models even on fully supervised language pairs (0.880 vs. 0.876 for FSEC), despite the latter being specifically optimized for these directions. The benefits become more pronounced for partially supervised (0.812 vs. 0.448 for BASE) and unsupervised (0.739 vs. 0.253 for BASE), although this improvement is expected as FS+PS+UN is explicitly fine-tuned on these directions.

These results clarify conflicting evidence on language diversity (see §1) and align with Wang et al.

[2] Preliminary experiments with additional FLORES-200 (dev) data showed no significant improvements, so we exclude it for experimental clarity.

[3] We use version wmt22-comet-da.

[4] We use the lid.176.bin model.

(2022) and Dang et al. (2024), confirming that *broad language diversity* (132 directions vs. 10–30), rather than minimal exposure, substantially enhances cross-lingual transfer, even for pairs already well supported in more specialized models.

Increased diversity reduces off-target problem Off-target translations, where models generate content in incorrect languages, represent a critical failure mode in LLM-based MT (Zhang et al., 2023; Guerreiro et al., 2023; Sennrich et al., 2024).

Figure 1 (right) shows that while all models maintain target language fidelity for fully supervised pairs, the BASE model produces incorrect target languages at alarming rates for partially supervised (44%) and unsupervised pairs (65%). Fine-tuning progressively mitigates this problem, with FS showing substantial improvements (13% and 31% respectively) despite not being explicitly trained on these language combinations. Significantly, the FS+PS+UN model completely eliminates off-target translations across all categories.

Diversity benefits plateau Expanding from FS+PS+UN (132 directions) to 272 directions reveals nuances in the diversity-performance relationship. Unsupervised directions benefit from increased diversity, while fully supervised directions show slight performance decline, suggesting benefits plateau beyond a certain threshold (details in Appendix C.1). This contradicts prior work that found monotonic improvements with diversity Wang et al. (2022); Dang et al. (2024), but aligns with Muennighoff et al. (2023)'s observation of diminishing returns when scaling multilingual pretraining beyond certain language counts.

Regularization alone insufficient Regularization benefits models by enhancing generalization and calibration, with strong effects when using distant languages (Meng and Monz, 2024). We investigate whether these benefits can be achieved through explicit regularization techniques (weight decay and LoRA) rather than language diversity, but find no comparable improvements. This aligns with Aharoni et al. (2019), who suggest that multilingualism provides benefits beyond explicit regularization methods. See Appendix C.2 for details.

Results not due to multi-parallel data While recent work by Caswell et al. (2025) found that fine-tuning on multi-parallel data causes catastrophic forgetting in LLMs when trained on X→en directions, our findings persist beyond multi-parallel settings. We replicated our experiments using non-multi-parallel data scraped from OPUS and observed similar diversity benefits (see Appendix C.3). Unlike the overfitting issues reported for LLMs, our models maintain performance, consistent with prior work showing multi-parallel data benefits in NMT (Stap et al., 2023; Wu et al., 2024).

Findings persist at larger scale Larger models (13B) exhibit the same trends: increased language diversity leads to reduced off-target rates and improved cross-lingual transfer. This confirms our findings are robust across model scales. Complete experimental details are provided in Appendix C.4.

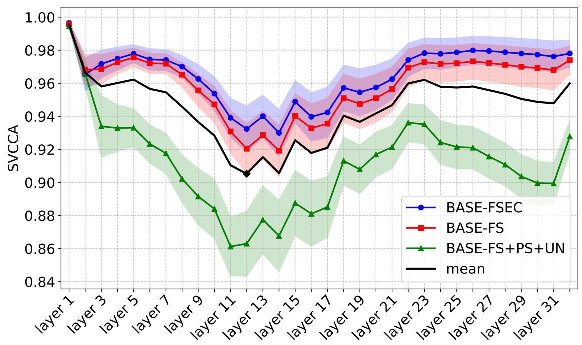

Figure 2: SVCCA similarity scores between fine-tuned and BASE models across layers. Lower values indicate greater adaptation during fine-tuning. BASE-FSEC (blue), BASE-FS (red), and BASE-FS+PS+UN (green) are compared, with their mean shown in black. Shaded regions represent confidence intervals. Middle layers show most significant adaptation, with lowest mean similarity (0.91) at layer 12. FP+PS+UN exhibits greater adaptation throughout the network.

Middle layers adapt most We analyze activation patterns across models by comparing them with the base model using Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al., 2017). This analysis identifies *where* and *to what extent* adaptations occur during fine-tuning. We aggregate activations across all source-target language pairs and present the layer-specific results in Figure 2.

Our analysis reveals that middle layers consistently undergo the most substantial adaptation across all fine-tuned models, with the lowest mean similarity (0.91) occurring at layer 12. Furthermore, models fine-tuned on more languages exhibit greater divergence from the base model, with FS+PS+UN showing most substantial adaptations.

Middle layers encode semantic information and show the strongest cross-lingual transfer capabilities (Liu and Niehues, 2025; Liu et al., 2025).

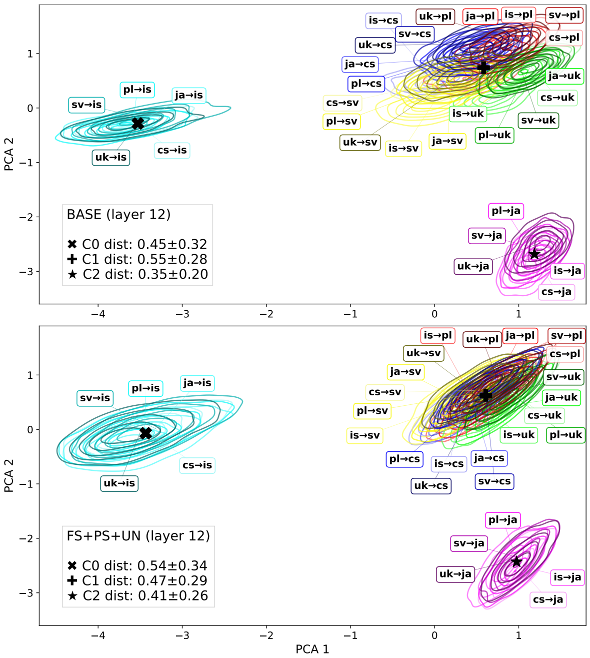

Figure 3: Kernel density estimation of layer 12 activations for BASE (top) and FS+PS+UN (bottom). Colors represent translation directions. Intra-cluster distances show increased specialization for single-target clusters in FS+PS+UN, while multi-target cluster C1 demonstrates increased overlap.

Our findings support that larger degrees of cross-lingual transfer within middle layers explain the performance improvements observed in models fine-tuned on a larger linguistic diversity.

Diversity increases cross-lingual overlap We analyze layer 12 (the most significantly modified layer) to understand *which adaptations* occur during fine-tuning. Following from Gao et al. (2024) and Wang et al. (2024), we apply t-SNE dimension reduction (van der Maaten and Hinton, 2008) to layer activations and visualize the bivariate kernel density (KDE) estimation. Next, we employ k-means clustering to identify language groups within these representations, using silhouette score maximization (Rousseeuw, 1987) for optimal cluster determination without requiring manual inspection. Finally, we calculate the intra-cluster distances. We compare the BASE and FS+PS+UN models, visualizing unsupervised directions where we expect the most significant adaptations.

Figure 3 presents the resulting visualization. Notably, for the single-target language clusters C0 and C2, the FS+PS+UN model exhibits *greater intra-cluster distances* (0.54±0.34 and 0.41±0.26) compared to the BASE model (0.45±0.32 and 0.35±0.20), suggesting *increased specialization* per source-target direction after fine-tuning on diverse data. Conversely, for the multi-target language cluster (C1), the FS+PS+UN model shows *reduced* intra-cluster distances (0.47±0.29) relative to the BASE model (0.55±0.28), indicating *greater representational overlap* between these linguistically related languages. This increased overlap provides evidence for enhanced cross-lingual transfer, which contributes to the superior performance of models fine-tuned on greater linguistic diversity.

Table 1 presents intra-cluster distances for all models. Note that clusters contain the same languages for all setups. As diversity increases, single-target clusters (C0, C2) show greater specialization while multi-language cluster C1 exhibits enhanced representational overlap, suggesting improved cross-lingual transfer.

While previous work has *explicitly* aligned representations (Liu and Niehues, 2025; Kargaran et al., 2024; Stap et al., 2023), our findings show *implicit* alignment occurs through multilingual fine-tuning.

	× C0	+ C1	⋆ C2
BASE	0.45 ± 0.32	0.55 ± 0.28	0.35 ± 0.20
FSEC	0.49 ± 0.33	0.53 ± 0.26	0.34 ± 0.20
FS	0.52 ± 0.36	0.51 ± 0.28	0.39 ± 0.24
FS+PS+UN	0.54 ± 0.34	0.47 ± 0.29	0.41 ± 0.26

Table 1: Intra-cluster distances. C0 (`is` target) and C2 (`ja` target) show increased distances in models fine-tuned on more diverse data, while C1 (`cs`, `pl`, `sv`, `uk` targets) shows decreased distances, indicating enhanced cross-lingual transfer.

4 Conclusion

Our systematic investigation across 132 translation directions resolves conflicting findings on language diversity in LLM fine-tuning. We show that fine-tuning on broader language sets consistently improves translation across all categories: fully supervised, partially supervised, and zero-shot pairs. Consequently, we recommend fine-tuning with diverse language directions even when optimizing for a limited subset of target translation pairs, as our most diverse model outperformed models specialized exclusively for those target pairs. However, we advise identifying an optimal diversity threshold, as too many languages diminishes performance for well-supported pairs while still benefiting less-represented languages. Our representational analysis attributes the diversity improvements to specific adaptations in middle layers, revealing increased language-agnostic representations, which explains the enhanced cross-lingual transfer.

Limitations

We evaluate on the FLORES-200 (Team et al., 2022) `devtest` set, a multi-parallel benchmark consisting of documents originally written in English and professionally translated into multiple languages. While this may introduce some translationese effects, the multi-parallel nature enables controlled comparison across language pairs.

Our findings are based on the TOWER model family (Alves et al., 2024) (7B and 13B), built on LLAMA 2 (Touvron et al., 2023). Further research should verify whether these patterns generalize to other model architectures and even larger model sizes.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First conference on language modeling*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. ArXiv:2502.12301 [cs].

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. ArXiv:2502.12404 [cs].

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – News Test References for MT Evaluation of 128 Languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards Boosting Many-to-Many Multilingual Machine Translation with Large Language Models. ArXiv:2401.05861 [cs].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der

Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai

Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517. Place: Cambridge, MA Publisher: MIT Press.

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 *International conference on learning representations*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. ArXiv:1612.03651 [cs].

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2024. MEXA: Multilingual Evaluation of English-Centric LLMs via Cross-Lingual Alignment. ArXiv:2410.05873 [cs].

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the ninth conference on machine translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2024. Neural methods for aligning large-scale parallel corpora from the web for south and East Asian languages. In *Proceedings of the ninth conference on machine translation*, pages 1454–1466, Miami, Florida, USA. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592. Place: Cambridge, MA Publisher: MIT Press.

Danni Liu and Jan Niehues. 2025. Middle-Layer Representation Alignment for Cross-Lingual Transfer in Fine-Tuned LLMs. ArXiv:2502.14830 [cs].

Weihao Liu, Ning Wu, Wenbiao Ding, Shining Liang, Ming Gong, and Dongmei Zhang. 2025. Selected languages are all you need for cross-lingual truthfulness transfer. In *Proceedings of the 31st international conference on computational linguistics*, pages 8963–8978, Abu Dhabi, UAE. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International conference on learning representations*.

Yan Meng and Christof Monz. 2024. Disentangling the roles of target-side transfer and regularization in multilingual machine translation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)*, pages 1828–1840, St. Julian's, Malta. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 15991–16111, Toronto, Canada. 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*, volume 30, pages 6076–6085. Curran Associates, Inc.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Aquia Richburg and Marine Carpuat. 2024. How multilingual are large language models fine-tuned for translation? In *First conference on language modeling*.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 2: Short papers)*, pages 21–33, St. Julian's, Malta. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 6189–6206, Bangkok, Thailand. Association for Computational Linguistics.

David Stap, Vlad Niculae, and Christof Monz. 2023. Viewing Knowledge Transfer in Multilingual Machine Translation Through a Representational Lens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14973–14987, Singapore. 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. ArXiv:2207.04672 [cs].

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2024. Bridging the Language Gaps in Large Language Models with Inference-Time Cross-Lingual Intervention. ArXiv:2410.12462 [cs].

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? Unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the association for computational linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2025. Contrastive preference optimization: pushing the boundaries of LLM performance in machine translation. In *Proceedings of the 41st international conference on machine learning*, ICML'24. JMLR.org.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19488–19496. Abstract note: Open-sourced large language models (LLMs) have demonstrated remarkable efficacy in various tasks with instruction tuning. However, these models can sometimes struggle with tasks that require more specialized knowledge such as translation. One possible reason for such deficiency is that instruction tuning aims to generate fluent and coherent text that continues from a given instruction without being constrained by any task-specific requirements. Moreover, it can be more challenging to tune smaller LLMs with lower-quality training data. To address this issue, we propose a novel framework using examples in comparison to teach LLMs to learn translation. Our approach involves output comparison and preference comparison, presenting the model with carefully designed examples of correct and incorrect translations and an additional preference loss for better regularization. Empirical evaluation on four language directions of WMT2022 and FLORES-200 benchmarks shows the superiority of our proposed method over existing methods. Our findings offer a new perspective on fine-tuning LLMs for translation tasks and provide a promising solution for generating high-quality translations. Please refer to Github for more details: https://github.com/lemon0830/TIM.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th international conference on machine learning*, ICML'23. JMLR.org. Place: Honolulu, Hawaii, USA Number of pages: 19 tex.articleno: 1722.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024a. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024b. A preference-driven paradigm for enhanced translation with large language models. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pages 3385–3403, Mexico City, Mexico. Association for Computational Linguistics.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the ninth conference on machine translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A Language details

The selection of languages shown in Table 2, following the language selection from Richburg and Carpuat (2024), enables evaluation across varied typological properties and scripts while providing a systematic comparison between supervised languages (seen during fine-tuning) and zero-shot languages that share linguistic features with the supervised set. The languages in the zero-shot set were chosen to represent both varying degrees of resource support in the pre-training data and to have relationships to languages in the supervised set through language family, typological properties, or orthography.

B Implementation details

B.1 Optimization

We conducted hyperparameter tuning on our development set (FLORES-200 `dev`), exploring learning rate scheduler $\in$ {cosine, inverse square root}, batch size $\in \{128, 256\}$, and learning rate $\in \{2 \times 10^{-5}, 2 \times 10^{-6}\}$.

For all experiments, we performed full fine-tuning using the AdamW optimizer (Loshchilov and Hutter, 2019) with 5% warm-up percentage and trained for one epoch. Based on development set performance, we selected the optimal configuration: a cosine learning rate scheduler with batch size of 256 and learning rate of 2×10^{-5}. We implemented our fine-tuning experiments using the Hugging Face transformers library (Wolf et al., 2020) with DeepSpeed (Rasley et al., 2020).

B.2 Inference

For both fine-tuning and zero-shot inference, we used the prompt template shown in Table 3. We mask out the prompt during fine-tuning. We employed greedy decoding (beam size 1) to balance computational efficiency with comprehensive evaluation across all translation directions.

Language	ISO 639-1	Script	LLaMA-2 support	Similarity groups
Czech	cs	Latin	0.03%	West Slavic
Polish	pl	Latin	0.09%	West Slavic
Russian	ru	Cyrillic	0.13%	East Slavic
Ukrainian	uk	Cyrillic	0.07%	East Slavic
German	de	Latin	0.17%	West Germanic
English	en	Latin	89.70%	West Germanic
Icelandic	is	Latin	possibly unseen	North Germanic
Dutch	nl	Latin	0.12%	West Germanic
Swedish	sv	Latin	0.15%	North Germanic
Japanese	ja	Kana	0.10%	Kanji from Hanzi, SOV order
Korean	ko	Hangul	0.06%	SOV order
Chinese	zh	Hanzi	0.13%	Hanzi to Kanji, loanwords to ja and ko

Table 2: Evaluated languages with rationales for similarity grouping, following the language selection from Richburg and Carpuat (2024). Languages marked in **bold** belong to the supervised set used in the original TOWER model fine-tuning.

```
Translate this from {source_language} to {target_language}:
{source_language}:  {source_sentence}
{target_language}:  {target_sentence}
```

Table 3: Prompting template for fine-tuning and 0-shot inference. For fine-tuning {target_sentence} is filled with the corresponding target sentence, and for 0-shot inference it is the empty string.

C Additional results

C.1 Scaling diversity to 272 languages

To investigate whether further increasing language diversity yields additional benefits, we compared our most diverse model from the main experiments (FS+PS+UN with 132 directions) to an even more diverse setup including 272 translation directions. While maintaining a similar distribution of language families as in our main experiments, we added five additional languages:

- **Germanic family**: Danish (da, North Germanic) and Afrikaans (af, West Germanic)
- **Slavic family**: Slovak (sk, West Slavic) and Bulgarian (bg, South Slavic)
- **East Asian languages**: Vietnamese (vi, different writing system but shares vocabulary with Chinese)

This selection maintains balanced representation across language families while introducing controlled diversity within each family. All additional languages are represented in both NTREX and FLORES-200.

Importantly, we evaluate both models on the same set of languages and directions as used throughout the paper. The additional languages are only used during fine-tuning to increase diversity, allowing us to measure their impact on the original set of translation directions.

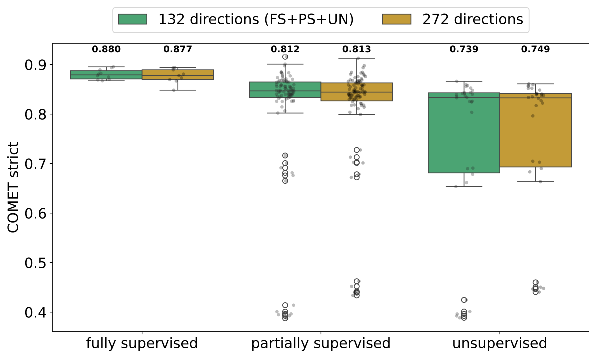

Figure 4: COMET-STRICT scores comparing models trained on 132 directions and 272 directions. Both are evaluated on the original test set with the same language pairs as used throughout the paper. Unsupervised directions show clearest benefits from increased diversity (+0.01), while fully supervised directions show a slight decrease (-0.003), suggesting potential diversity trade-offs.

Figures 4 shows the performance comparison between our 132-direction model (FS+PS+UN) and the expanded 272-direction model.

For fully supervised pairs, we observe a slight performance decrease (-0.003 COMET-STRICT) when scaling to 272 directions. Partially supervised directions show almost identical performance (+0.001), while unsupervised directions demonstrate the clearest benefit (+0.01) from increased language diversity.

These results suggest that language diversity benefits may plateau or even slightly decline for already well-represented language pairs. The slight reduction in fully supervised performance may indicate a trade-off between focused optimization and broader generalization, where extremely high diversity can dilute the model's effectiveness for specific well-represented languages. Nonetheless, the continued improvements for unsupervised directions (with respect to the original TowerBase model) demonstrate that higher diversity provides additional benefits for these previously unseen language combinations, even though both the 132 and 272 direction models include these pairs during fine-tuning.

C.2 Regularization alone insufficient

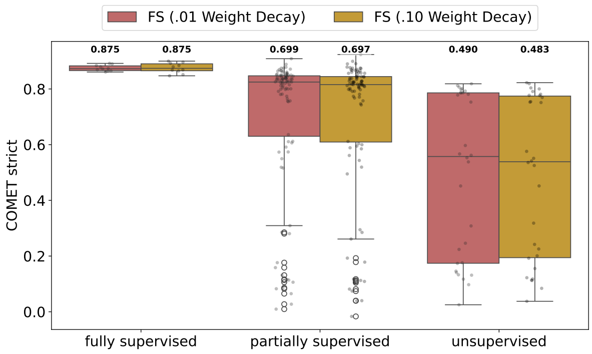

Figure 5: COMET-STRICT scores comparing FS models with weight decay values of 0.01 (standard) and 0.10. Increasing regularization strength shows minimal impact on fully supervised and partially supervised directions, while actually harming performance on unsupervised directions, suggesting that regularization alone cannot replicate the benefits of increased language diversity.

To investigate whether the performance benefits observed with increased language diversity could be achieved through explicit regularization techniques, we conduct additional experiments using stronger regularization on models with limited language diversity. If increased language diversity primarily functions as a form of regularization, we hypothesize that similar improvements could be obtained by directly increasing regularization strength in less diverse models.

All our previous experiments use the AdamW optimizer with weight decay set to 0.01 and gradient clipping at 1.0. This aligns with common practices in LLM fine-tuning, where dropout (Srivastava et al., 2014) is rarely employed (neither the Llama nor Tower papers mention dropout, though both use weight decay). Notably, AdamW applies weight decay directly to the weights rather than through gradients, decoupling it from the learning rate.

We tested this hypothesis by fine-tuning the FS setup with increased weight decay values of 0.05 and 0.10 (compared to our standard 0.01). We chose the FS setup to examine whether stronger regularization would induce better cross-lingual transfer to partially supervised and unsupervised directions, potentially mimicking the benefits observed in the more diverse FS+PS+UN model.

Figure 5 shows the COMET-STRICT scores comparing FS models with weight decay values of 0.01 (standard) and 0.10.[5] Increasing the regularization strength has minimal impact on translation performance across all language categories. For fully supervised directions, both models achieved identical mean scores (0.875). For partially supervised directions, the difference was negligible (0.699 vs. 0.697). For unsupervised directions, the model with stronger regularization actually performed slightly worse (0.483 vs. 0.490).

We further explored alternative regularization approaches by implementing LoRA (Hu et al., 2022) with rank 64, which constrains fine-tuning to a low-dimensional subspace. This parameter-efficient tuning method can be considered a form of regularization as it restricts model updates to a much smaller parameter space than full fine-tuning, potentially preventing overfitting. Results from LoRA experiments align with our weight decay findings: performance for fully and partially supervised directions remained comparable to full fine-tuning with standard regularization, while unsupervised directions showed slight degradation.

These experiments demonstrate that our initial weight decay value of 0.01 already provides an appropriate balance between overfitting prevention and model flexibility. More importantly, they confirm that the cross-lingual transfer benefits observed in more diverse models cannot be replicated merely by increasing explicit regularization in less diverse models. The language diversity benefits we observe go beyond simple explicit regularization effects, providing specialized cross-lingual knowledge transfer. Our findings align with Aharoni et al. (2019), who suggest that multilingualism provides

[5]The results for weight decay at 0.05 were very similar to 0.10 and are omitted for clarity.

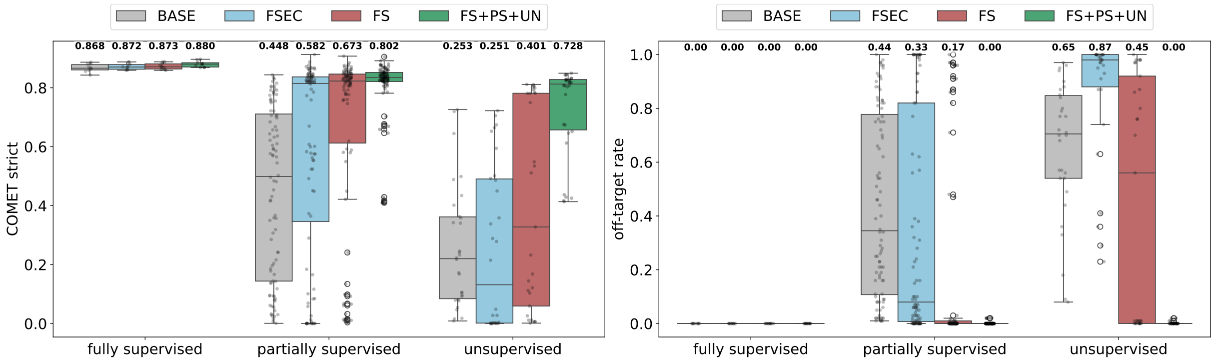

Figure 6: COMET-STRICT scores for 7B models fine-tuned on filtered NLLB dataset: BASE (no fine-tuning), FSEC (English-centric), FS (seen directions), FS+PS+UN (all directions), evaluated on *fully supervised* (de/en/ko/nl/ru/zh pairs), *unsupervised* (cs/is/ja/pl/sv/uk pairs), and *partially supervised* (combining supervised and unsupervised) language pairs. Numbers above bars show mean scores. Training on more diverse sets improves *all* categories, with FS+PS+UN achieving best results even for fully supervised pairs. FS substantially reduces off-target rates for unsupervised directions compared to BASE and FSEC, despite these pairs being *absent* from its fine-tuning data.

benefits beyond what can be achieved through explicit regularization methods.

C.3 Results not due to multi-parallel data

To verify our findings are not artifacts of using multi-parallel data, we constructed a non-multi-parallel dataset from the NLLB corpus (Team et al., 2022). We maintained the same 132 language directions as in our main experiments but eliminated the multi-parallel property Following Koehn (2024), we extract examples with LASER (Artetxe and Schwenk, 2019) scores above 1.05. We then removed sentences that appeared in multiple language pairs and sampled the remaining data to ensure exactly 2,000 examples per direction, creating a completely non-multi-parallel dataset of equivalent size to our NTREX experiments.

Figure 6 shows COMET-STRICT (left) and off-target (right) results from experiments conducted using the filtered NLLB dataset rather than NTREX, allowing us to verify that our findings are not artifacts of using multi-parallel data.

The results demonstrate that our core finding—increased language diversity during fine-tuning leads to better performance—holds when using non-multi-parallel data as well. The FS+PS+UN model still achieves the highest COMET-STRICT scores across all language categories, including for fully supervised language pairs. This confirms that the benefits of diverse fine-tuning extend beyond the multi-parallel setting described in our main experiments.

When comparing performance between models fine-tuned on NLLB versus NTREX data, we observe identical ranking patterns across different fine-tuning setups, though the NTREX-trained models show slightly better overall performance. This marginal improvement is likely attributable to NTREX's higher data quality, as it consists of professionally translated content specifically designed for machine translation evaluation.

C.4 Invariance to model scale

Figure 7 demonstrates that our findings about language diversity benefits persist when scaling to 13B parameters.

For translation quality (Figure 7, left), the most diverse setup (FS+PS+UN) consistently achieves the best results across all language categories, including fully supervised pairs. While most 13B models show higher scores than their 7B counterparts (Figure 1, left), the FSEC model unexpectedly performs worse than BASE in partially supervised and unsupervised settings (0.557 vs 0.558 and 0.154 vs 0.465), unlike in the 7B configuration where FSEC outperformed BASE.

For off-target rates (Figure 7, right), the most diverse setup again eliminates off-target translations completely. No model produces off-target translations for fully supervised pairs. The FSEC 13B model shows substantially worse performance for partially supervised (0.34) and unsupervised (0.80) pairs compared to its 7B version (Figure 1, right). Though BASE and FS 13B models show improved off-target rates compared to 7B, the problem remains significant (BASE: 39% for unsupervised, FS: 22%).

The decrease in performance for the FSEC 13B model can likely be attributed to overfitting to the limited English-centric training data.

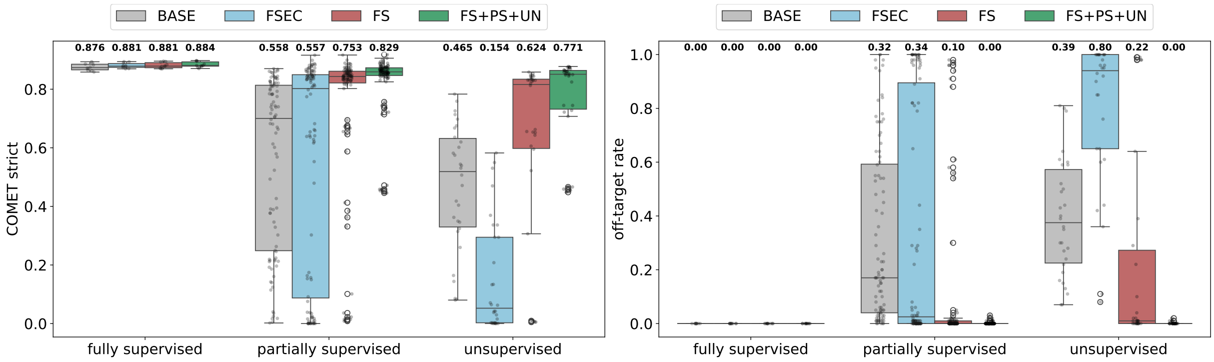

Figure 7: COMET-STRICT scores for 13B models: BASE (no fine-tuning), FSEC (English-centric), FS (seen directions), FS+PS+UN (all directions), evaluated on *fully supervised* (de/en/ko/nl/ru/zh pairs), *unsupervised* (cs/is/ja/pl/sv/uk pairs), and *partially supervised* (combining supervised and unsupervised) language pairs. Numbers above bars show mean scores. Training on more diverse sets improves *all* categories, with FS+PS+UN achieving best results even for fully supervised pairs. FS substantially reduces off-target rates for unsupervised directions compared to BASE and FSEC, despite these pairs being *absent* from its fine-tuning data.

These results confirm that language diversity benefits during fine-tuning are robust across model scales, consistently improving both translation quality and target language fidelity.