Cross-lingual Transfer Dynamics in BLOOMZ: Insights into Multilingual Generalization

Sabyasachi Samantaray and Preethi Jyothi
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{sachiray, pjyothi}@cse.iitb.ac.in

Abstract

Multilingual large language models have emerged as a promising solution for resource-constrained settings, with significant efforts aimed towards improving multilingual capabilities of English-centric pretrained models. However, the broader cross-lingual implications of fine-tuning interventions remain understudied. This work examines instruction tuning (IT) over the BLOOMZ model for Question Answering (QA) in low-resource settings, with special emphasis on transfer dynamics across several languages. Our findings reveal two critical insights: first, IT on the target language can negatively impact its own performance in constrained short-span generation tasks due to overgeneration tendencies; second, in QA tasks, IT appears to suppress performance in some interfering languages, thereby enhancing capabilities in some target Indic languages by *more than doubling* QA performance. These results highlight important trade-offs in multilingual LLM adaptation and enhance our understanding of cross-lingual transfer mechanisms.

1 Introduction

Large language models (LLMs) excel in tasks like classification, text generation, and information extraction. Recently, cross-lingual alignment has been widely studied to enhance the multilingual capabilities of LLMs (Zhu et al., 2024; Zhang et al., 2024; Hu et al., 2021). Since most of the world's languages can be deemed low-resource owing to the limited amounts of high-quality data (Asai et al., 2024; Razumovskaia et al., 2024), cross-lingual alignment is an important problem to tackle.

Prior work on multilinguality has largely focused on cross-lingual dynamics within English-centric models and pretrained decoder-only models (Zhao et al., 2024; Xu et al., 2023; Wendler et al., 2024). In this work, we study the cross-lingual abilities of BLOOMZ, a multilingual, multi-task instruction-tuned model (Muennighoff et al., 2023; Scao et al., 2022). We focus on a constrained generation task, closed question answering (QA), that can be objectively evaluated (unlike open-ended generation tasks like machine translation) while still being vulnerable to generation-related artefacts (unlike classification tasks). We examine the impact of instruction tuning (IT) on QA on several Indic and non-Indic languages. Surprisingly, we find significant performance improvements using languages that transcend language family relatedness and surface-level script similarities, indicating that BLOOMZ exhibits cross-lingual generalization beyond typological proximity (Ifergan et al., 2024). We also present a new multilingual logit lens-based analysis to provide more insights into cross-lingual dynamics that result in performance improvements or degradations. Our analysis reveals two key phenomena: 1) Suppression of the target language and 2) a tendency to over-generate in the target language, both of which significantly affect the model's output as illustrated in Figure 1.

Alignment with prior work. A growing consensus from recent work (Zhao et al., 2024; Wendler et al., 2024) is that English-centric models like Llama (Touvron et al., 2023) and Mistral (Jiang et al., 2023) process multilingual prompts by internally mapping to an English "thinking space" in intermediate layers, while the initial and final layers are multilingual in nature. Concurrent work explores language-specific neurons in the multilingual LLMs and reveals their significant roles in the outer layers (Tang et al., 2024; Zhu et al., 2024). Kargaran et al. (2024) argues that stronger alignment between English and non-English middle layer embeddings correlate with better cross-lingual transfer.

While current LLMs and adaptation methods demonstrate promising surface-level cross-lingual

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 47–61

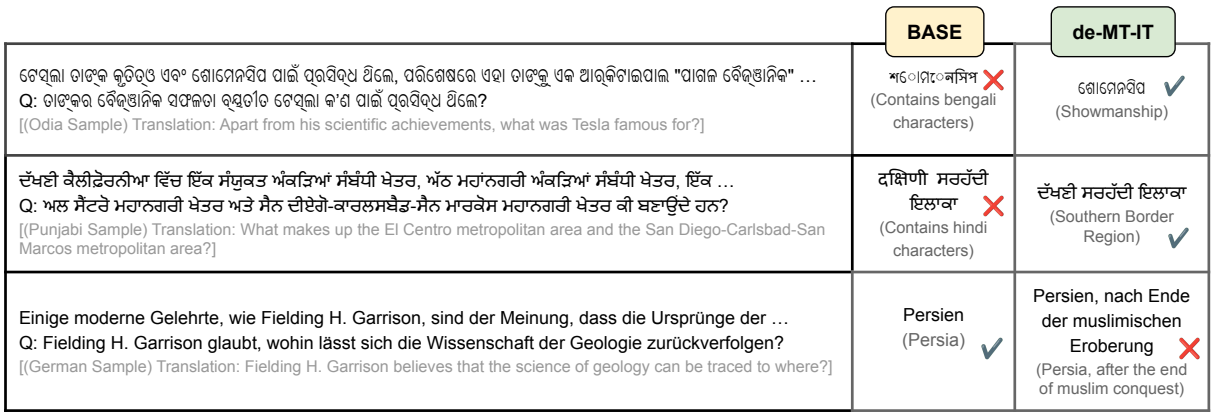

Figure 1: Examples demostrating that Instruction tuning on a small German-QA train set (generated via NLLB-MT) improves Odia and Punjabi performance by suppressing interference from Bengali and Hindi, respectively, but leads to overgeneration on German. Complete passages omitted for brevity.

abilities (on tasks like style transfer), they struggle with deeper cross-lingual reasoning and knowledge transfer. This limitation suggests the presence of a cross-lingual knowledge barrier, as noted by Chua et al. (2024). Towards addressing this gap, it has been observed that fine-tuning on certain languages can improve the performance of others, indicating the presence of cross-lingual bridging mechanisms (Singh et al., 2024b; Bai et al., 2024; Ifergan et al., 2024; Wang et al., 2024; Bai et al., 2023). Our experiments also support this possibility of cross-lingual bridging mechanisms. Our findings align with Ifergan et al. (2024) who documented BLOOM's unique ability to facilitate factual recall across languages with different scripts.

2 Methodology

2.1 Logit Lens

Understanding how knowledge propagates through the layers of a model is critical for gaining insights into the internal workings of multilingual LLMs. One such interpretive tool is the Logit Lens, introduced by nostalgebraist (2020). This technique provides a mechanism to probe the latent representations in intermediate layers by mapping them directly to vocabulary probabilities using the last layer's linear language modeling head. In prior work, Zhao et al. (2024) used logit lens to investigate the multilingual alignment of intermediate representations in Vicuna-13B-v1.5 (Chiang et al., 2023) and BLOOMZ-7B1 (Muennighoff et al., 2023). Similarly, Wendler et al. (2024) utilized logit lens to analyze intermediate representations in Llama models (Touvron et al., 2023) to measure the token probabilities for English and Chinese words across different layers.

2.2 Probing for Language Identification

In our work, we adopt the logit lens framework to examine the flow of linguistic knowledge across the layers of BLOOMZ-7B1 for a diverse set of languages[1]. Our analysis leverages Cook and Lui's (2012) `langid.py` script, which assigns a probability distribution over languages for each token in the model's vocabulary. Tokens composed solely of punctuation or numeric digits (0–9) are excluded, as they do not belong to a specific language and add noise to the analysis. The `langid.py` tool supports 96 languages, covering all languages in our experiments. We compute per-layer language probabilities by multiplying the per-token language probabilities with the token probabilities obtained from layer embeddings transformed via the language modeling head. To ensure statistical robustness, this process is repeated across multiple test set samples, and the final latent probabilities are derived by averaging the language distributions across all samples. Equation 1 estimates probability of language L at a layer j, given a dataset of task-specific examples D and a vocabulary V.

$$P_j(L) = \frac{1}{|D|} \sum_{D_i \in D} \sum_{t \in V} P_j(t|D_i) P(L|t) \quad (1)$$

Further details about formatting of the question and context in each task-specific example is given in Appendix A.

[1]Code and dataset is available at https://github.com/Sachi-27/Multilingual-NLP.

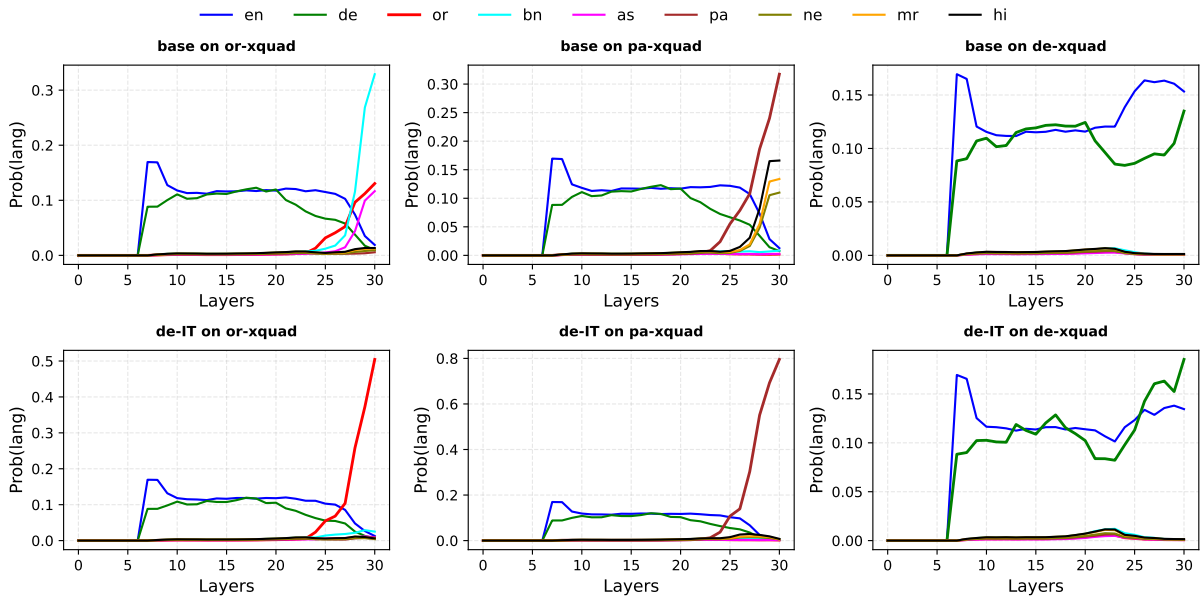

Figure 2: Comparison of logit lens plots for the BLOOMZ base model before and after German IT (de-IT i.e.,instruction tuning on the German train set), evaluated on Odia (or), Punjabi (pa) and German (de) xquad test sets. For or-xquad and pa-xquad, the plots consider samples corrected after de-IT. In contrast, for de-xquad, the plots consider samples degraded after de-IT. Whether the response is corrected or degraded is decided based on the Exact Match Metric.

Unlike naive token frequency analyses, this method gives a probabilistic measure of language dominance, allowing us to capture subtle shifts in multilingual representation.

2.3 Experimental Setup

We conduct our experiments using the BLOOMZ-7B1[2] model and the multilingually parallel dataset, IndicGenBench's XQuAD-IN (Singh et al., 2024a), which comprises data in 12 Indic languages and English. From this dataset, we utilize 103 context-question-answer triplets for training for 10 epochs. To construct a parallel training set for non-Indic languages, we leverage translations from NLLB (Costa-jussà et al., 2022) and Opus-MT (or MarianMT) (Tiedemann et al., 2023). Non-Indic test sets are sourced from the XQuAD dataset in the TensorFlow dataset library (Artetxe et al., 2019). Additionally, for robustness, we incorporate IndicQA (Doddapaneni et al., 2023), an out-of-domain question-answering dataset in Indic languages[3]. All experiments are evaluated in a zero-shot setting. The metrics are reported using Token-F1 and Exact Match scores (Rajpurkar et al., 2016). More details are in Appendix B.

3 Results and Discussion

Suppression for Performance Improvements. We highlight Token-F1 scores for 5 Indic languages: Gujarati (gu), Kannada (kn), Malayalam (ml), Odia (or), and Punjabi (pa) as these showed significant improvements in performance. Metrics are presented in Table 1, with the detailed results for all languages is available in Appendix D. Notably, with just 103 Russian-translated samples, we observe significant improvements, particularly a doubling of performance for the low-resource language Odia.

Through analysis of the multilingual logit lens plots (Figure 2), we identify distinct "hill-like" patterns in the middle-layer latents of languages such as German (de), Estonian (et), Swedish (sv), Xhosa (xh), Finnish (fi), Indonesian (id), and Malay (ms). We verify that, to some extent, these languages can facilitate cross-lingual transfer, in accordance with the findings of Zhao et al. (2024). We compare the logit lens plots for or-xquad and pa-xquad test sets between the base model and the de-IT model (i.e., the base model finetuned on German train set). We focus on samples where the base model answers wrongly, but the de-IT model provided correct predictions, shown in Figure 2. These plots highlight

[2]Choice is constrained by the model's multilingual nature and QA-specific instruction tuning.

[3]Other datasets like TyDiQA and MLQA are leaked into BLOOMZ and thus unsuitable for evaluation.

Method	gu	kn	ml	or	pa
Base	**60.81**	**48.52**	**49.07**	**25.90**	**55.80**
en-IT	56.05	50.96	49.26	31.65	63.56
gu-IT	43.68	48.06	44.73	43.30	53.47
kn-IT	58.76	48.87	50.03	51.68	66.01
ml-IT	58.88	53.74	43.21	**52.56**	68.88
or-IT	50.32	48.12	45.49	48.11	61.67
pa-IT	53.78	50.06	48.03	48.34	58.63
de-MT-IT	63.90	57.83	52.66	49.86	70.99
et-MT-IT	**66.18**	56.17	53.06	38.29	68.41
fi-MT-IT	64.48	57.10	**55.35**	43.78	70.42
ru-MT-IT	59.75	55.66	51.16	**56.38**	68.32
sv-MT-IT	65.06	**58.27**	53.72	50.96	70.28
th-MT-IT	65.15	**58.92**	**55.09**	51.22	**71.80**
tr-MT-IT	**65.91**	56.89	54.20	48.38	71.57
xh-MT-IT	62.93	57.94	54.90	49.27	**71.34**

Table 1: Token-F1 scores of Instruction Tuned (IT) models evaluated on XQuAD-IN test set. Languages denoted with "MT" indicate training data generated via NLLB machine translation from the English train set. Performance rankings are visually indicated with **green** (highest) and **blue** (second highest).

Model	or (XQuAD-IN)	or (IndicQA)
Base	25.90, 17.05	26.82, 13.68
or-IT	48.41, 30.08	35.64, 18.37
de-MT-IT	49.86, 32.94	46.14, 27.91
sv-MT-IT	50.96, 33.52	43.61, 26.11

Table 2: Performance metrics (Token-F1 score, Exact Match Score) of Instruction-Tuned models evaluated on Odia test sets from XQuAD-IN and IndicQA.

token suppression in related languages, such as the reduction in Bengali and Assamese latent probabilities in the final layers for Odia, and the similar suppression of Hindi, Marathi, and Nepali for Punjabi. This in turn results in an increase in the last-layer probabilities for the target languages – Odia and Punjabi, in this case – that correlates with performance improvements. Similar trends are also observed across other languages, as detailed in Appendix C.

We also conduct out-of-domain evaluations using the IndicQA Odia datasets, comparing performance on or-xquad with different languages for instruction tuning. Our findings reveal similar performance gains, as shown in Table 2.

High-Resource Fine-Tuning is Not Universally Beneficial. Contrary to prior work, our findings challenge the notion that fine-tuning on high-resource languages universally improves performance across the multilingual spectrum. The performance metrics of en-IT model on gu-xquad test set serves as a clear example of this. Surprisingly, certain medium to low-resource languages, such as Kannada, Malayalam, Thai and Turkish contribute significantly to overall model improvement. This suggests that the effectiveness of fine-tuning languages in enhancing alignment and generalization is not solely dependent on data availability.

Self-performance trade-off. Self-IT (i.e., IT using language X evaluated on test samples of language X) appears to negatively impact performance on QA-style tasks that require concise, span-based answers. This is likely due to the model's tendency to generate verbose (and sometimes hallucinatory) responses, that negatively affects task accuracy. Logit lens plots in Figure 2 illustrate this tendency for de-IT on de-xquad, with rising latent probabilities in the final layers indicative of over-generation.

Better Translation Quality Leads to Improved Cross-Lingual Transfer. Results in Tables 3 and 11 are consistent. For example, in Indonesian, NLLB generated translations are of better quality, correlating with better performance on IT. Logit lens visualizations (Figure 8) shows that IT with MarianMT translations struggles with Bengali suppression, while IT with NLLB translations enables Odia to surpass Bengali, aligning with id-IT gains. More details are in Appendix D.

IT Lang	MT Model	BLEU	gu	kn	ml	or	pa
Base	-	-	**60.81**	**48.52**	**49.07**	**25.90**	**55.80**
id	MarianMT	19.99	51.13	44.46	34.40	21.96	55.57
	NLLB	46.97	54.52	48.52	37.52	32.95	61.00
xh	MarianMT	8.42	59.36	55.22	53.34	46.18	68.88
	NLLB	23.88	62.93	57.94	54.90	49.27	71.34

Table 3: Comparison of IT models trained on machine translated training data using MarianMT vs NLLB and evaluated on XQuAD-IN test set. Here, green highlights the higher Token-F1 scores and blue highlights MT with higher BLEU scores.

4 Conclusion

This work highlights the intricate nature of multilingual task specific fine-tuning and its diverse effects across languages. We demonstrate that instruction tuning with a very small set of samples is unlikely to acquire substantial new knowledge, but can induce shifts in linguistic structures, particularly in the later layers, leading to suppression of interfering language latents, contributing to improved

performance. However, these improvements are neither uniform nor guaranteed, as high-resource fine-tuning does not always yield positive effects, and self-IT performance trade-offs often emerge. Moreover, the quality of training data significantly influences outcomes, with better translation quality directly correlating with improved multilingual alignment.

5 Limitations

Our study highlights the significance of latent structures in the intermediate layers of multilingual models, emphasizing their role in cross-lingual transferability. We also demonstrate that instruction tuning impacts performance across languages differently, influenced by their intrinsic characteristics and resource levels. However, our analysis is subject to several limitations. Our experiments focus only on the task of span-based question answering and one specific multilingual model, BLOOMZ. Other multilingual models such as Gemma-7B (Team et al., 2024) and Aya-13B (Üstün et al., 2024) exhibit very irregular and unstructured logit lens plots. Their plots deviate from the multilingual hypothesis (Zhao et al., 2024), which posits that multilingual models predominantly "think" in English or Latin-centric representations. Instead, these models exhibit a significant mix-up in thinking across languages and deviate from "hill" type latent representations, indicating a different latent structure than what is typically observed in conventional Latin-centric multilingual models. This restricts the generalizability of our findings to other models and task types. Finally, although we observe that languages like German (de) and Swedish (sv) trigger suppression to improve performance, the underlying mechanism behind this phenomenon remains unclear, warranting further investigation.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. 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 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yang Bai, Anthony Colas, Christan Grant, and Zhe Wang. 2024. M3: A multi-task mixed-objective learning framework for open-domain multi-hop dense sentence retrieval. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10846–10857, Torino, Italia. ELRA and ICCL.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2024. Crosslingual capabilities and knowledge barriers in multilingual large language models. *arXiv preprint arXiv:2406.16135*.

Paul Cook and Marco Lui. 2012. langid.py for better language modelling. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 107–112, Dunedin, New Zealand.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Hai Hu, He Zhou, Zuoyu Tian, Yiwen Zhang, Yina Patterson, Yanting Li, Yixin Nie, and Kyle Richardson. 2021. Investigating transfer learning in multilingual pre-trained language models through Chinese natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3770–3785, Online. Association for Computational Linguistics.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2024. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. *arXiv preprint arXiv:2408.10646*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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 preprint arXiv:2410.05873*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

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.

nostalgebraist. 2020. Interpreting gpt: the logit lens. LessWrong. Retrieved from https://www.lesswrong.com/posts/8Q4QpK7F8F8G/interpreting-gpt-the-logit-lens.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2024. Analyzing and adapting large language models for few-shot multilingual nlu: Are we there yet? *arXiv preprint arXiv:2403.01929*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024a. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. 2024b. A three-pronged approach to cross-lingual adaptation with multilingual llms. *arXiv preprint arXiv:2406.17377*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8051, Miami, Florida, USA. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Multilingual Logit Lens Implementation Details

The language confidences scores for each token are obtained from `langid.py` and stored in a token-language probability table, $\mathcal{T}$, where $\mathcal{T}_{ij}$ represents the probability of token t_i belonging to language j. Mathematically, this can be represented as:

$$\mathcal{T}_{tl} = \sigma\left(c_{\text{langid}}(t, l)\right) \tag{2}$$

where $c_{\text{langid}}(t, l)$ denotes the confidence score output by `langid.py` for token t belonging to each language l, and σ is the softmax function, ensuring that the probabilities across all languages sum to 1 for every token. We exlude tokens consisting of only numbers and punctuation marks, by zeroing all entries in the table corresponding to such a token t. For BLOOMZ-7B1, there are 6,269 such tokens out of 250,680 in its vocabulary.

For each input sample, the logit lens is applied to the embeddings $h_n^{(j)}$ at every layer j of the model for the last input token x_n. The logits obtained from these embeddings are then transformed into language probabilities by mapping them with the token-language probability table $\mathcal{T}$. This mapping is expressed as:

$$P(\text{lang} = l \mid h_n^{(j)}) = \sum_t \mathcal{T}_{tl} \cdot \sigma(\text{logit}(h_n^{(j)}))[t] \tag{3}$$

where $\sigma(\text{logit}(h_n^{(j)}))$ represents the logits of the embedding $h_n^{(j)}$ obtained after passing the embedding through the linear modelling head. This operation provides a distribution over languages for the embeddings at every layer j. Additionally, to address a specific model behavior, we implement probability zeroing for tokens corresponding to "A:" in the initial layers. This post-processing step is necessary because the model exhibits a tendency to overly weight "A:" tokens, due to their presence as the final token in the input prompt (Table 4).

[Context in Target Language] **Q:** [Question in Target Language] **A:**

Table 4: Standardized prompt template for Question Answering, aligned with the format used in IndicGenBench.

B Experimental Setup Details

B.1 Datasets

The dataset splits used in our experiments are reported in Table 5. IndicGenBench's XQuAD-IN consists of English (en) and 12 Indic languages: Assamese (as), Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Odia (or), Punjabi (pa), Tamil (ta), Telugu (te), and Urdu (ur). The test set comprises 1,190 examples and is fully parallel with the XQuAD dataset from the TensorFlow Datasets (TFDS) library. For training, XQuAD-IN includes a subset of 103 examples from the English training set of TFDS-XQuAD. To ensure parallelism and maintain consistency across instruction tuning (IT) experiments, we use this 103-example subset and translate it into other languages for fine tuning.

While the TensorFlow Datasets (TFDS) version of XQuAD contains training data for German (de), Russian (ru), Thai (th), Turkish (tr), Greek (el), Spanish (es), and Vietnamese (vi), our goal was to ensure a small parallel dataset across languages to fairly investigate cross-lingual effects. There are two key reasons why we opted for additional translations: 1, Avoiding bias from dataset discrepancies and 2, Consistency in machine translation sources.

To evaluate generalization, we also incorporate IndicQA, an out-of-domain question-answering test dataset covering the same 11 Indic languages as

Dataset	Train	Val	Test
XQuAD-IN	103	111	1190
TFDS-XQuAD	~80K	~10K	1190
IndicQA	-	-	~1K

Table 5: Dataset Splits

XQuAD-IN, excluding Urdu. Unlike XQuAD-IN, IndicQA is not parallel across languages.

B.2 Instruction Tuning Details

We utilized prompts tailored to the QA task as in Singh et al. (2024a). The causal language model (LM) is fine-tuned for 10 epochs using the PEFT LoRA framework (Mangrulkar et al., 2022), with updates restricted to the query-key-value layers of BLOOMZ. The fine-tuning follows a causal LM objective, maximizing the likelihood of generating the next token given the previous tokens. The process optimizes the generative probability of the complete prompt, which includes the context, question, and the correct response as shown in 4. Key hyperparameters for fine-tuning include a learning rate of 2×10^{-4}, LoRA rank $r = 64$, and $\alpha = 16$. All experiments were conducted on a single NVIDIA A100-SXM4-80GB GPU, with a max runtime of 3 minutes for 103 samples over 10 epochs.

$$\arg\min_{\phi} \sum_{\mathcal{P}=\{\mathcal{C},\mathcal{Q},\mathcal{R}\}\in\mathcal{D}} -\log p_{\phi}(\mathcal{P}) \qquad (4)$$

B.3 Evaluation

We conducted evaluations on the XQuAD-IN and TFDS test sets in a zero-shot setting. The outputs generated by the LLM are compared with the reference answers using the widely adopted SQuAD evaluation metrics (Rajpurkar et al., 2016). This reports the Token-level F1 score, which measures the overlap between predicted and ground-truth tokens, considering partial matches and the exact match (EM) score, which measures the strict match between the predicted answer and the reference. We used evaluations on the base BLOOMZ-7B1 model as the baseline. Our results are based on a single run, which is reproducible by setting random seeds. The generation process follows controlled decoding with top-k sampling ($k = 50$), nucleus sampling (top-p= 0.95), and generating a single output sequence (num_return_sequences=1) at temperature= 0.1.

C Logit Lens Plots

Figures 3 and 4 display the logit lens plots for Odia and Punjabi samples where both the base and de-IT models make the same correct predictions. Similar, albeit less pronounced, suppressions are observed for Gujarati, Malayalam, and Kannada in XQuAD-IN, leading to minimal performance gains (Figure 5). Comparable trends are observed for other IT languages, as detailed in Appendix Figures 6 and 7.

ISO Code	Language	ISO Code	Language
as	Assamese	bn	Bengali
de	German	el	Greek
en	English	es	Spanish
et	Estonian	fi	Finnish
fr	French	hi	Hindi
id	Indonesian	kn	Kannada
ml	Malayalam	mr	Marathi
ms	Malay	or	Odia
pa	Punjabi	ru	Russian
sv	Swedish	ta	Tamil
te	Telugu	th	Thai
tr	Turkish	vi	Vietnamese
xh	Xhosa		

Table 6: ISO Code to Language Mapping

D Complete Metrics

We report the complete performance metrics on the XQuAD-IN test set for its 12 Indic languages in Tables 7 and 8. Additionally, we present scores on the TFDS XQuAD test sets for 7 languages in Table 9. Malay (ms) is excluded from our analysis because NLLB doesn't support it. We report the IndicQA test performance for 6 languages across several selected IT languages in Table 10. Furthermore, we present complete performance metrics of IT models trained on 6 middle-layer hill languages (de, et, fi, id, sv, and xh). The training data are machine translated from English. Performance comparisons of MarianMT and NLLB generated train data are provided in Tables 11 and 12. To measure translation quality, we use BLEU scores for the training contexts, comparing MT-generated outputs against Google Translate generations (used as ground truth).

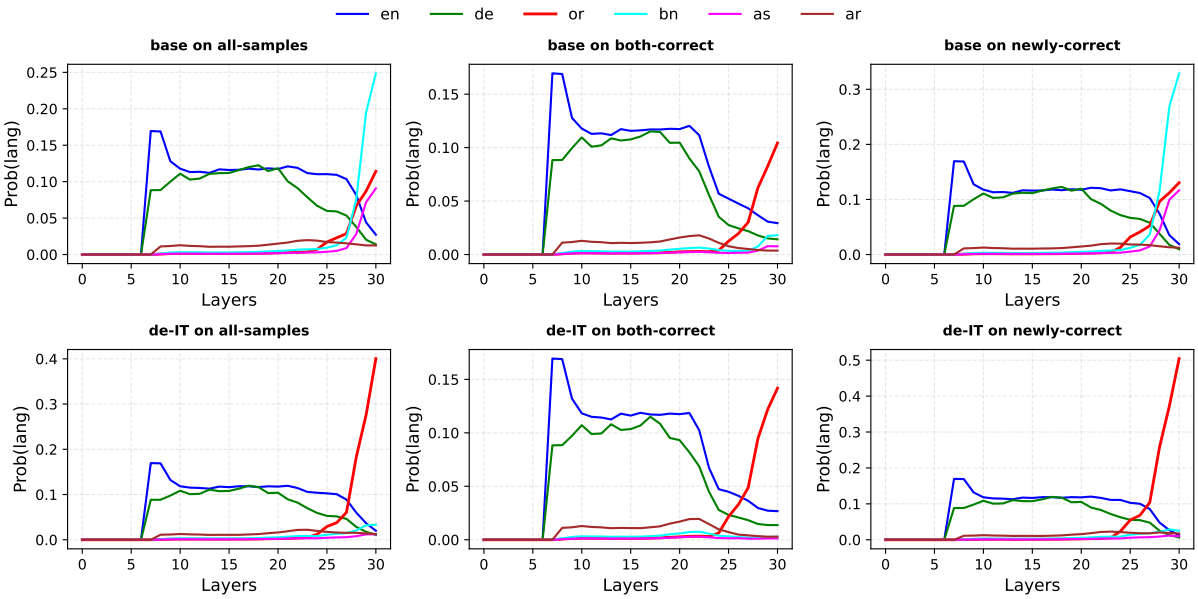

Figure 3: Logit lens analysis of the BLOOMZ model before and after German IT (de-IT) on Odia test data (or-xquad). The plots illustrate three scenarios considering: all samples, samples with correct predictions across both models, and newly corrected samples—those misclassified by the base model but correctly predicted after de-IT. Samples with correct predictions on base model have low interference. Correction of predictions occur where Bengali (bn) and other interference (Assamese (as) and Arabic (ar)) is suppressed and replaced by stronger Odia (or) signals.

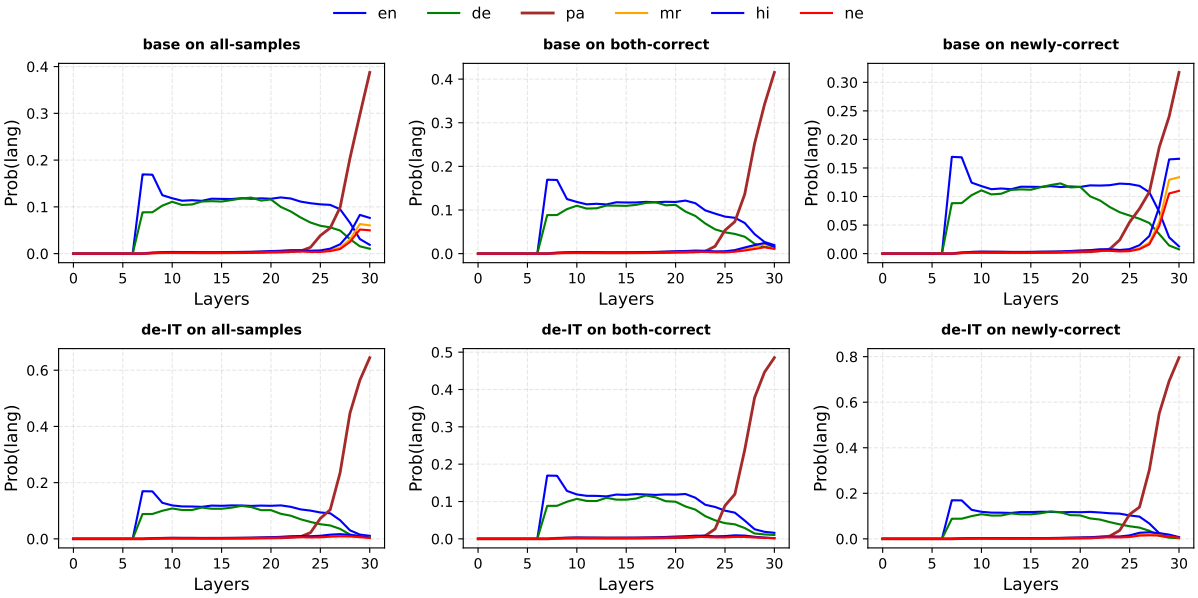

Figure 4: Logit lens analysis of the BLOOMZ model before and after German IT (de-IT) on Punjabi test data (pa-xquad). The plots illustrate three scenarios considering: all samples, samples with correct predictions across both models, and newly corrected samples—those misclassified by the base model but correctly predicted after de-IT. Samples with correct predictions on base model have low interference. Correction of predictions occur when interfering latents of Hindi (hi), Marathi (mr) and Nepali (ne) are suppressed and replaced by stronger Punjabi (pa) signals.

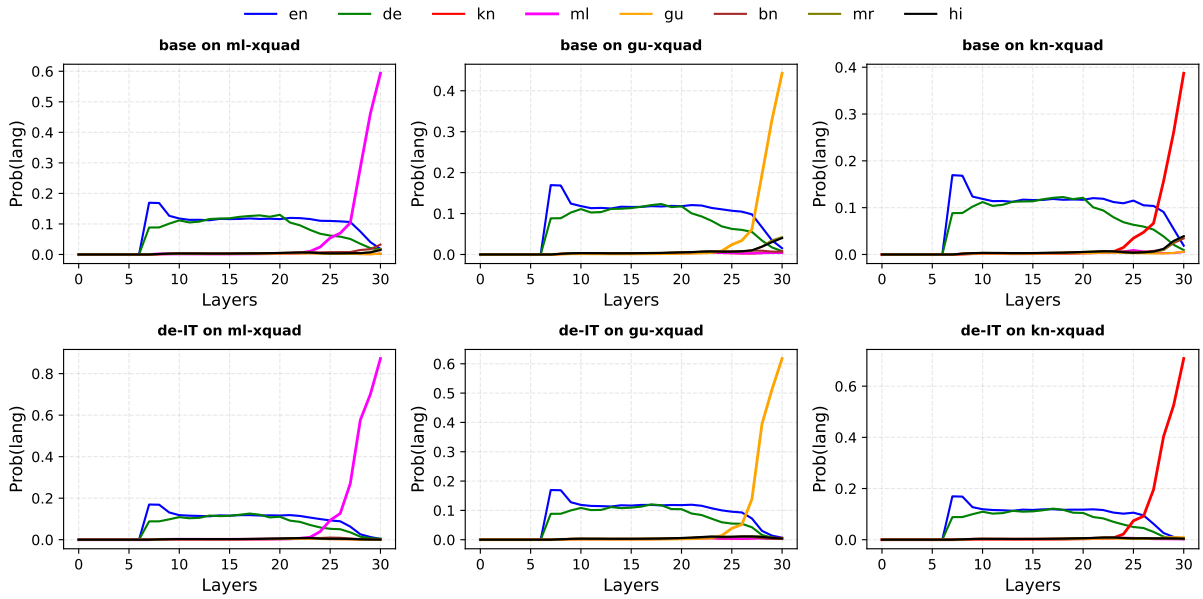

Figure 5: Comparison of logit lens plots for the BLOOMZ model before and after German IT (de-IT), evaluated on test data from Malayalam (ml), Gujarati (gu) and Kannada (kn) xquad test sets. The plots consider samples misclassified by the base model but correctly predicted after de-IT. There is a rise in test language probability in the last layers after de-IT indicating stronger signals correlating with improved performance.

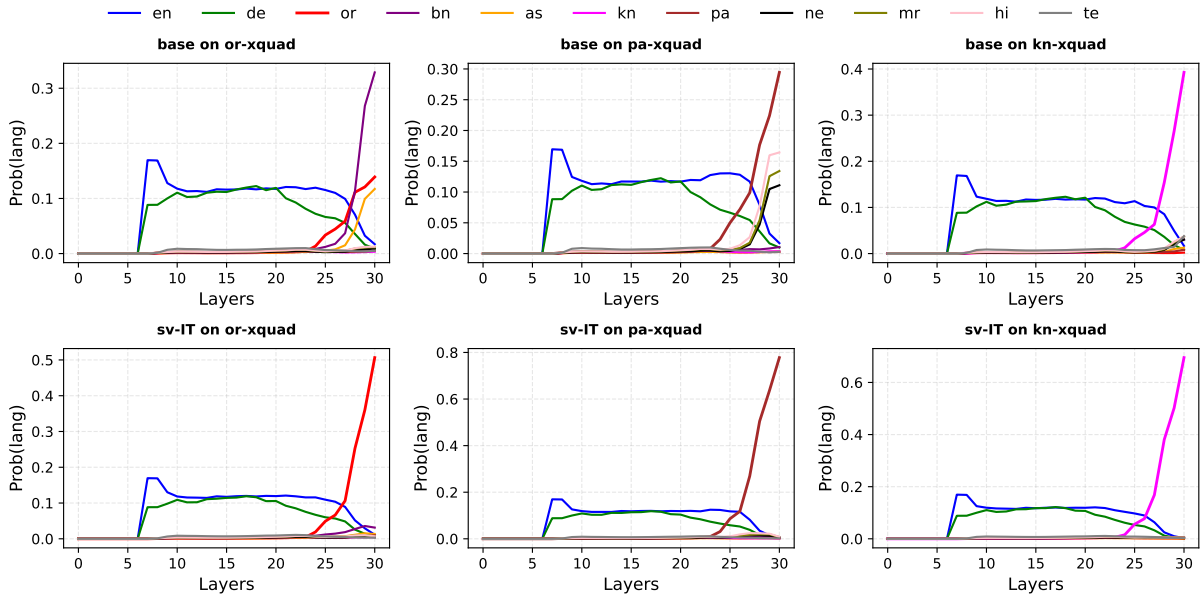

Figure 6: Comparison of logit lens plots for the BLOOMZ model before and after Swedish IT (sv-IT), evaluated on test data from Odia (or), Punjabi (pa) and Kannada (kn) xquad test sets. The plots consider samples misclassified by the base model but correctly predicted after sv-IT. There is a rise in Kannada latent in the last layers after sv-IT indicating stronger signals correlating with improved performance. Suppression of interfering languages in Odia and Punjabi after sv-IT correlates with improved performance.

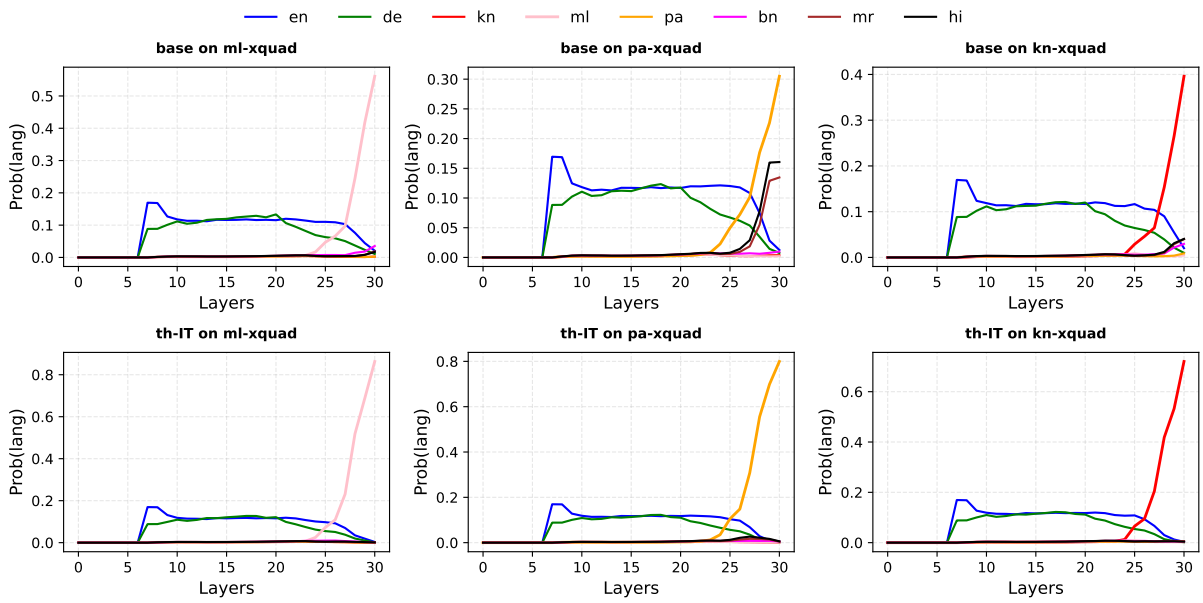

Figure 7: Comparison of logit lens plots for the BLOOMZ model before and after Thai IT (th-IT), evaluated on test data from Malayalam (ml), Punjabi (pa) and Kannada (kn) xquad test sets. The plots consider samples misclassified by the base model but correctly predicted after th-IT. Suppression of interfering last layer latents in Punjabi and rising last layer signals in Malayalam and Kannada correlate with improved performance.

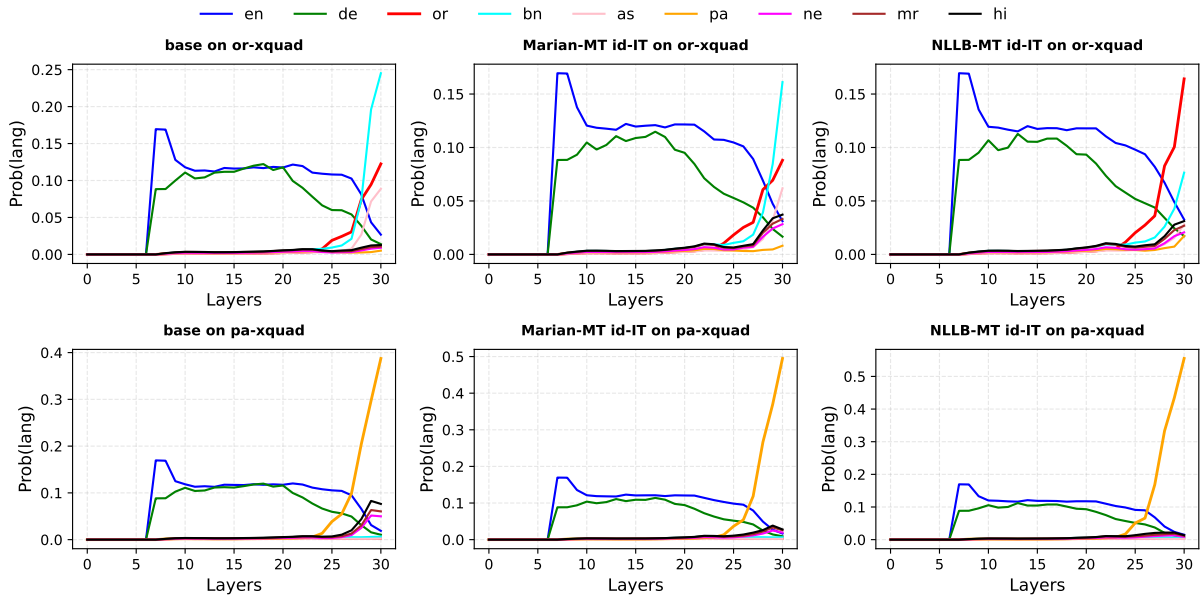

Figure 8: Logit lens comparison of the BLOOMZ model before and after Indonesian IT (id-IT), utilizing training data derived from the English parallel subset of IndicGenBench with machine translations from MarianMT (Helsinki-Opus) and NLLB. The analysis is conducted on all test samples from Odia (or) and Punjabi (pa) xquad test sets.

Token F1, EM	gu	kn	ml	or	pa	ur
Base	**60.81, 43.94**	**48.52, 32.35**	**49.07, 33.44**	**25.90, 17.05**	**55.80, 38.73**	**67.75, 49.07**
as-IT	55.41, 35.96	53.20, 34.95	46.77, 31.09	46.28, 28.82	67.10, 46.80	65.30, 46.80
bn-IT	49.60, 30.08	43.47, 27.56	43.52, 27.64	23.55, 14.53	58.35, 34.28	59.13, 37.64
en-IT	56.05, 37.14	50.96, 33.19	49.26, 31.93	31.65, 20.08	63.56, 43.10	62.71, 41.42
gu-IT	43.68, 24.20	48.06, 30.33	44.73, 27.89	43.30, 25.71	53.47, 28.82	62.74, 42.26
hi-IT	48.14, 28.06	47.83, 31.34	45.19, 29.32	31.86, 19.07	58.09, 33.44	58.44, 34.78
kn-IT	58.76, 41.17	48.87, 30.36	50.03, 34.28	51.68, 34.28	66.01, 46.80	65.56, 46.47
ml-IT	58.88, 40.00	53.74, 35.37	43.21, 25.71	**52.56, 35.88**	68.88, 49.66	65.97, 46.97
mr-IT	55.11, 37.31	51.36, 34.11	48.16, 32.35	49.42, 32.68	63.49, 44.53	64.44, 45.79
or-IT	50.32, 31.68	48.12, 31.59	45.49, 30.25	48.11, 30.25	61.67, 40.08	64.00, 44.87
pa-IT	53.78, 36.05	50.06, 33.19	48.03, 32.77	48.34, 32.43	58.63, 37.56	61.83, 40.00
ta-IT	50.09, 30.25	49.02, 30.33	47.11, 30.92	45.02, 26.80	63.36, 38.99	63.86, 43.19
te-IT	51.30, 31.93	48.34, 31.09	44.89, 29.24	40.19, 24.78	63.73, 41.17	64.15, 45.12
ur-IT	51.71, 33.44	47.55, 30.75	43.94, 27.73	34.37, 21.93	57.88, 35.79	56.70, 33.10
de-MT-IT	63.90, 45.96	57.83, 40.00	52.66, 36.47	49.86, 32.94	70.99, 52.10	66.92, 47.14
el-MT-IT	64.05, 45.54	**58.35, 39.66**	**55.37, 37.89**	51.22, 34.20	70.44, 50.08	66.96, 47.73
es-MT-IT	47.87, 29.41	45.70, 28.99	40.85, 25.46	27.60, 17.05	54.73, 32.18	53.29, 30.92
et-MT-IT	**66.18, 48.73**	56.17, 38.57	53.06, 37.39	38.29, 25.54	68.41, 49.66	67.04, 47.56
fi-MT-IT	64.48, 47.22	57.10, 39.24	**55.35, 39.41**	43.78, 28.9	70.42, 51.84	66.98, 47.47
fr-MT-IT	48.87, 30.16	46.50, 27.98	39.51, 24.70	36.42, 22.01	52.47, 27.64	51.09, 28.90
id-MT-IT	54.52, 33.94	48.52, 31.51	37.52, 23.94	32.95, 21.42	61.00, 39.83	59.64, 38.57
ru-MT-IT	59.75, 40.75	55.66, 35.63	51.16, 33.94	**56.38, 38.40**	68.32, 46.97	65.88, 45.71
sv-MT-IT	65.06, 47.39	58.27, 40.58	53.72, 37.14	50.96, 33.52	70.28, 51.59	**67.83, 47.98**
th-MT-IT	65.15, 46.63	**58.92, 40.84**	55.09, 39.57	51.22, 33.94	**71.80, 52.77**	68.01, 48.40
tr-MT-IT	**65.91, 48.73**	56.89, 39.07	54.20, 37.98	48.38, 32.35	**71.57, 32.35**	67.68, 48.57
vi-MT-IT	50.41, 31.59	45.16, 28.40	38.10, 25.54	28.50, 18.15	55.50, 31.68	56.66, 35.12
xh-MT-IT	62.93, 43.27	57.94, 40.16	54.90, 38.31	49.27, 31.59	71.34, 51.93	**67.97, 48.31**

Table 7: Performance metrics (Token-F1 score, Exact Match score) of Instruction Tuned (IT) models on the XQuAD-IN test set. Training data sourced from the XQuAD-IN parallel corpus. Languages denoted with "MT" indicate training data generated via NLLB machine translation from the English subset. Performance rankings are visually indicated with **green** (highest) and **blue** (second highest) based on Token F1 scores. The results are shown for languages gu (Gujarati), kn (Kannada), ml (Malayalam), or (Odia) and ur (Urdu).

Method	bn	te	hi	mr	as	ta
Base	**64.20, 44.53**	**58.22, 39.74**	**73.02, 48.82**	**65.96, 50.00**	**51.33, 32.43**	**63.37, 45.88**
as-IT	60.19, 38.31	55.86, 36.89	72.09, 47.31	58.76, 40.50	45.26, 24.03	58.61, 40.75
bn-IT	54.93, 32.85	52.26, 33.94	67.23, 38.15	52.54, 34.03	40.80, 22.18	52.67, 33.27
en-IT	60.58, 39.15	55.11, 35.96	69.79, 42.43	60.61, 43.86	49.42, 28.90	53.04, 34.62
gu-IT	60.91, 39.41	51.55, 32.94	70.38, 44.70	51.79, 33.36	46.60, 27.31	56.48, 38.82
hi-IT	56.86, 33.19	52.32, 34.45	61.60, 31.42	52.83, 31.68	46.50, 27.14	54.12, 35.21
kn-IT	63.33, 42.35	54.98, 37.05	70.84, 46.80	61.72, 44.36	51.19, 31.93	59.53, 41.76
ml-IT	63.32, 41.84	56.18, 38.31	72.31, 47.73	62.36, 46.05	51.01, 31.59	57.91, 41.34
mr-IT	61.75, 40.67	53.32, 35.56	69.95, 45.04	52.19, 34.20	46.75, 28.31	59.51, 42.52
or-IT	58.76, 39.24	50.86, 32.52	70.07, 46.13	55.58, 38.90	22.11, 11.34	59.73, 40.76
pa-IT	61.05, 39.83	54.06, 36.30	69.45, 44.20	56.51, 39.74	47.75, 28.15	57.64, 40.75
ta-IT	60.71, 38.23	49.48, 29.83	70.92, 45.37	56.18, 37.98	49.46, 29.83	46.76, 26.38
te-IT	62.31, 41.76	45.84, 28.23	70.41, 45.79	59.06, 42.10	47.56, 27.98	52.01, 33.86
ur-IT	58.65, 37.22	52.46, 34.45	68.56, 42.01	56.52, 37.14	46.19, 26.97	53.77, 35.63
de-MT-IT	65.30, 44.28	57.80, 39.57	73.29, 47.64	64.47, 47.89	52.80, 32.26	61.20, 43.86
el-MT-IT	**65.59, 44.28**	57.67, 38.82	**73.67, 48.48**	65.80, 48.57	53.02, 32.52	60.93, 43.78
es-MT-IT	53.47, 31.68	51.18, 31.76	62.66, 34.36	49.76, 33.69	43.15, 22.94	47.89, 27.39
et-MT-IT	65.55, 45.29	**58.88, 41.26**	73.55, 48.82	**66.77, 50.16**	53.50, 33.78	**62.73, 46.38**
fi-MT-IT	65.49, 44.87	**58.86, 40.84**	**73.65, 48.15**	66.20, 49.32	53.64, 33.69	62.21, 45.04
fr-MT-IT	54.60, 32.52	51.06, 32.60	58.73, 29.57	49.76, 30.67	44.92, 26.38	49.62, 31.17
id-MT-IT	56.83, 34.53	54.26, 34.28	67.32, 39.66	55.39, 38.31	47.32, 27.81	52.46, 33.27
ru-MT-IT	64.84, 42.60	58.70, 39.24	73.06, 46.89	64.08, 45.96	53.19, 32.26	58.10, 40.50
sv-MT-IT	64.90, 44.36	57.98, 39.49	73.64, 48.15	66.02, 49.91	53.54, 32.94	62.70, 46.30
th-MT-IT	65.42, 45.21	58.07, 39.91	73.44, 48.73	**66.38, 49.83**	**54.08, 33.36**	61.46, 44.95
tr-MT-IT	65.52, 45.21	58.67, 40.67	73.62, 48.90	65.62, 49.66	**54.31, 33.44**	62.29, 45.71
vi-MT-IT	55.44, 34.11	52.14, 34.20	66.03, 36.80	52.17, 35.71	44.93, 26.55	52.39, 32.52
xh-MT-IT	**65.72, 45.21**	58.54, 39.91	73.52, 49.15	64.31, 47.73	53.57, 32.43	62.06, 45.21

Table 8: Performance metrics (Token-F1 score, Exact Match score) of Instruction Tuned (IT) models on the XQuAD-IN test set. Training data sourced from the XQuAD-IN parallel corpus. Languages denoted with "MT" indicate training data generated via NLLB machine translation from the English subset. Performance rankings are visually indicated with **green** (highest) and **blue** (second highest) based on Token-F1 scores. The results are shown for languages bn (Bengali), te (Telugu), hi (Hindi), mr (Marathi), as (Assamese) and ta (Tamil).

METHOD	ru	el	es	de	vi	th	tr
Base	**61.95, 45.21**	**38.31, 28.31**	**89.15, 75.88**	**71.93, 57.05**	**88.07, 76.21**	**21.04, 16.21**	**34.47, 24.70**
as-IT	57.25, 41.59	34.12, 23.44	88.00, 72.60	67.38, 53.02	86.89, 73.44	19.95, 13.69	29.90, 20.16
bn-IT	55.40, 39.91	35.42, 23.52	85.18, 67.22	64.19, 48.99	84.32, 68.06	19.13, 13.94	21.75, 14.78
en-IT	57.30, 40.92	34.64, 22.26	72.04, 51.59	66.06, 49.49	78.28, 61.00	21.97, 15.63	29.33, 17.81
gu-IT	56.25, 39.07	33.23, 22.43	87.24, 71.68	66.80, 51.59	87.62, 74.53	19.60, 14.36	27.66, 19.15
hi-IT	57.58, 41.09	31.59, 19.66	83.90, 64.78	64.90, 50.42	85.07, 68.48	18.13, 12.43	23.42, 14.53
kn-IT	59.30, 42.60	35.30, 23.94	88.15, 73.69	68.34, 52.85	87.53, 74.28	**23.05, 16.55**	30.75, 22.26
ml-IT	56.54, 41.68	34.96, 23.60	88.50, 73.44	68.91, 55.21	87.35, 73.94	20.55, 15.29	29.30, 20.33
mr-IT	57.49, 41.68	34.21, 23.27	87.76, 71.68	66.98, 52.43	87.30, 74.20	20.93, 15.96	28.16, 18.82
or-IT	57.53, 41.84	32.95, 21.34	87.40, 72.01	67.46, 52.43	86.25, 72.52	18.77, 13.86	25.05, 17.89
pa-IT	56.79, 41.00	34.13, 22.26	86.75, 72.01	66.83, 51.68	86.22, 72.68	19.91, 13.69	27.08, 19.32
ta-IT	58.90, 42.18	**37.33, 25.63**	87.29, 73.02	69.08, 54.28	86.75, 73.02	**24.05, 17.56**	29.55, 18.65
te-IT	56.13, 40.58	35.73, 25.04	88.68, 74.53	66.79, 52.43	86.61, 73.36	21.06, 15.63	30.48, 21.00
ur-IT	53.19, 37.98	32.51, 21.34	86.35, 70.16	66.12, 51.09	85.39, 69.74	15.52, 11.59	23.77, 16.63
de-MT-IT	53.46, 36.80	31.29, 20.33	87.21, 72.18	44.27, 30.33	87.12, 74.11	16.33, 12.60	19.57, 13.61
el-MT-IT	55.06, 38.48	15.01, 3.86	88.28, 74.28	68.16, 51.68	**88.15, 75.88**	16.87, 11.17	29.29, 17.56
es-MT-IT	51.38, 35.79	27.28, 16.80	50.84, 28.40	56.55, 40.75	68.11, 46.97	15.12, 10.58	18.68, 11.09
et-MT-IT	57.85, 42.35	35.99, 24.45	88.67, 74.53	68.41, 53.27	87.54, 75.04	21.49, 15.71	19.83, 8.31
fi-MT-IT	59.13, 43.52	36.41, 25.12	88.61, 74.45	67.99, 52.35	87.83, 75.46	21.83, 15.12	24.57, 12.68
fr-MT-IT	50.89, 34.45	28.03, 17.14	61.69, 38.90	54.88, 38.15	65.28, 43.69	15.58, 11.34	17.81, 9.91
id-MT-IT	55.06, 37.89	35.02, 22.35	73.75, 51.00	63.90, 47.89	71.68, 48.40	20.23, 14.11	24.91, 15.54
ru-MT-IT	28.94, 12.10	20.71, 12.77	87.37, 71.68	64.80, 48.40	86.76, 73.02	7.67, 4.53	22.16, 15.04
sv-MT-IT	56.89, 41.26	36.57, 24.62	87.64, 72.68	63.10, 48.06	87.09, 73.94	19.67, 14.36	22.98, 14.36
th-MT-IT	**60.92, 45.37**	34.17, 21.93	89.03, 74.95	**71.79, 55.71**	87.98, 75.71	8.13, 1.42	**32.24, 22.52**
tr-MT-IT	58.71, 43.19	35.63, 23.78	**89.34, 75.46**	68.10, 52.94	87.87, 75.46	21.36, 15.29	16.20, 6.89
vi-MT-IT	51.83, 35.46	29.52, 19.15	74.54, 51.93	62.49, 46.89	67.23, 42.43	14.72, 10.75	20.49, 12.18
xh-MT-IT	59.90, 44.03	35.08, 24.20	88.24, 73.61	68.23, 52.85	87.67, 74.53	19.75, 13.94	22.37, 13.36

Table 9: Performance metrics (Token-F1 score, Exact Match score) of Instruction Tuned (IT) models on the TFDS-XQuAD test set. Training data sourced from the XQuAD-IN parallel corpus. Languages denoted with "MT" indicate training data generated via NLLB machine translation from the English subset. Performance rankings are visually indicated with **green** (highest) and **blue** (second highest) based on Token-F1 scores. The results are shown for languages ru (Russian), el (Greek), es (Spanish), de (German), vi (Vietnamese), th (Thai) and tr (Turkish).

Model	gu	hi	kn	ml	or	pa
Base	**56.88, 41.25**	**79.56, 64.54**	**44.65, 28.14**	**48.48, 31.66**	**26.82, 13.68**	**62.84,40. 47**
ml-IT	54.19, 35.74	78.15, 63.02	42.84, 25.32	38.88, 22.11	42.53, 25.09	60.87, 37.51
de-MT-IT	**58.94, 42.49**	79.47, 63.87	45.72, 27.96	48.32, 31.21	46.14, 27.91	**63.2, 40.13**
ru-MT-IT	58.02, 40.76	78.69, 62.35	**47.39, 29.11**	**49.65, 32.21**	**46.8, 28.61**	62.4, 39.11
sv-MT-IT	56.48, 39.51	79.21, 63.59	43.97, 26.56	48.32, 31.21	43.61, 26.11	62.22, 39.2

Table 10: Performance metrics (Token-F1 score, Exact Match score) of Instruction Tuned (IT) models on the out-of-domain IndicQA test set. Training data sourced from the XQuAD-IN parallel corpus. Languages denoted with "MT" indicate training data generated via NLLB machine translation from the English subset. Highest Performance ranking based on Token-F1 scores are visually indicated with **green**. The results are shown for languages gu (Gujarati), hi (Hindi), kn (Kannada), ml (Malayalam), or (Odia) and pa (Punjabi) over models instrution tuned on ml (Malayalam), de (German), ru (Russian) and sv (Swedish).

IT Lang	MT Model	BLEU	gu	kn	ml	or	pa
Base	-	-	**60.81, 43.94**	**48.52, 32.35**	**49.07, 33.44**	**25.90, 17.05**	**55.80, 38.73**
de	MarianMT	40.57	64.46, 45.54	57.83, 39.41	54.18, 37.98	52.03, 34.78	71.14, 51.00
	NLLB	34.13	63.90, 45.96	57.83, 40.00	52.66, 36.47	49.86, 32.94	70.99, 52.10
et	MarianMT	28.27	65.85, 48.65	58.26, 39.91	53.18, 37.31	36.64, 24.03	67.58, 49.41
	NLLB	24.33	66.18, 48.73	56.17, 38.57	53.06, 37.39	38.29, 25.54	68.41, 49.66
fi	MarianMT	23.96	65.15, 47.64	58.03, 39.24	54.83, 39.07	46.88, 31.09	69.94, 51.00
	NLLB	22.80	64.48, 47.22	57.10, 39.24	55.35, 39.41	43.78, 28.90	70.42, 51.84
id	MarianMT	19.99	51.13, 32.77	44.46, 27.98	34.40, 22.26	21.96, 14.21	55.57, 32.60
	NLLB	46.97	54.52, 33.94	48.52, 31.51	37.52, 23.94	32.95, 21.42	61.00, 39.83
sv	MarianMT	56.65	64.88, 46.38	58.14, 40.25	54.15, 37.22	49.50, 32.43	70.86, 51.51
	NLLB	41.35	65.06, 47.39	58.27, 40.58	53.72, 37.14	50.96, 33.52	70.28, 51.59
xh	MarianMT	8.42	59.36, 39.15	55.22, 36.72	53.34, 36.97	46.18, 30.33	68.88, 48.48
	NLLB	23.88	62.93, 43.27	57.94, 40.16	54.90, 38.31	49.27, 31.59	71.34, 51.93

Table 11: Comparison of MT models trained on machine translated training data using MarianMT (Helsinki-opus) vs NLLB for different IT languages and evaluated on selected test languages – gu (Gujarati), kn (Kannada), ml (Malayalam), or (Odia) and pa (Punjabi). Here, green is decided based on higher token-F1 scores and blue highlights the MT with higher BLEU score.

IT Lang	MT Model	en	bn	te	hi	mr	as	ta	ur
Base	-	**93.32, 85.79**	**64.20, 44.53**	**58.22, 39.74**	**73.02, 48.82**	**65.96, 50.00**	**51.33, 32.43**	**63.37, 45.88**	**67.75, 49.07**
de	MarianMT	91.81, 83.69	65.47, 44.70	58.28, 39.66	73.71, 48.40	65.06, 48.57	54.63, 34.11	61.81, 45.29	67.30, 47.64
	NLLB	91.76, 83.36	65.30, 44.28	57.80, 39.57	73.29, 47.64	64.47, 47.89	52.80, 32.26	61.20, 43.86	66.92, 47.14
et	MarianMT	93.45, 85.88	65.92, 45.71	58.20, 40.42	72.97, 48.40	65.29, 48.31	54.52, 33.69	62.60, 45.71	67.21, 45.76
	NLLB	92.93, 85.21	65.55, 45.29	58.88, 41.26	73.55, 48.82	66.77, 50.16	53.50, 33.78	62.73, 46.38	67.04, 47.56
fi	MarianMT	93.03, 85.63	65.10, 44.53	58.77, 40.75	73.46, 48.48	65.83, 50.33	52.54, 33.36	61.66, 44.70	66.75, 47.47
	NLLB	93.01, 85.63	65.49, 44.87	58.86, 40.84	73.65, 48.15	66.20, 49.32	53.64, 33.69	62.21, 45.04	66.98, 47.47
id	MarianMT	81.92, 66.38	54.50, 31.17	52.95, 33.52	63.95, 33.69	51.51, 32.85	43.40, 23.44	51.89, 30.58	58.66, 36.47
	NLLB	80.08, 65.96	56.83, 34.53	54.26, 34.28	67.32, 39.66	55.39, 38.31	47.32, 27.81	52.46, 33.27	59.64, 38.57
sv	MarianMT	92.43, 84.62	64.46, 43.61	58.04, 39.74	72.98, 47.64	65.48, 48.82	53.49, 32.60	61.35, 45.54	67.14, 47.89
	NLLB	92.16, 84.20	64.90, 44.36	57.98, 39.49	73.64, 48.15	66.02, 49.91	53.54, 32.94	62.70, 46.30	67.83, 47.98
xh	MarianMT	92.97, 85.12	65.14, 44.11	57.13, 38.90	73.06, 48.31	62.54, 44.87	50.61, 30.00	59.20, 42.10	66.23, 46.21
	NLLB	92.43, 84.78	65.72, 45.21	58.54, 39.91	73.52, 49.15	64.31, 47.73	53.57, 32.43	62.06, 45.21	67.97, 48.31

IT Lang	MT Model	ru	el	es	de	vi	th	tr
Base	-	**61.95, 45.21**	**38.31, 28.31**	**89.15, 75.88**	**71.93, 57.05**	**88.07, 76.21**	**21.04, 16.21**	**34.47, 24.70**
de	MarianMT	55.36, 38.48	34.48, 23.69	87.60, 72.35	49.14, 33.52	87.54, 74.28	20.38, 15.21	27.25, 18.82
	NLLB	53.46, 36.80	31.29, 20.33	87.21, 72.18	44.27, 30.33	87.12, 74.11	16.33, 12.60	19.57, 13.61
et	MarianMT	58.59, 43.27	36.77, 25.46	88.66, 74.78	68.10, 53.86	87.84, 75.12	22.15, 16.30	27.14, 14.70
	NLLB	57.85, 42.35	35.99, 24.45	88.67, 74.53	68.41, 53.27	87.54, 75.04	21.49, 15.71	19.83, 8.31
fi	MarianMT	58.00, 41.68	35.52, 24.03	88.68, 74.78	66.77, 50.75	87.43, 75.04	20.90, 14.70	23.07, 11.34
	NLLB	59.13, 43.52	36.41, 25.12	88.61, 74.45	67.99, 52.35	87.83, 75.46	21.83, 15.12	24.57, 12.68
id	MarianMT	51.24, 33.69	32.09, 19.83	67.66, 43.78	57.87, 41.17	68.75, 45.37	17.25, 12.43	19.62, 11.00
	NLLB	55.06, 37.89	35.02, 22.35	73.75, 51.00	63.90, 47.89	71.68, 48.40	20.23, 14.11	24.91, 15.54
sv	MarianMT	58.30, 41.84	36.01, 24.03	87.78, 73.10	62.59, 46.63	87.63, 74.53	15.90, 11.34	19.19, 12.43
	NLLB	56.89, 41.26	36.57, 24.62	87.64, 72.68	63.10, 48.06	87.09, 73.94	19.67, 14.36	22.98, 14.36
xh	MarianMT	56.38, 40.42	30.50, 20.84	87.98, 73.86	65.89, 51.42	87.55, 74.11	16.91, 12.43	20.40, 14.28
	NLLB	59.90, 44.03	35.08, 24.20	88.24, 73.61	68.23, 52.85	87.67, 74.53	19.75, 13.94	22.37, 13.36

Table 12: Comparison of MT models trained on machine translated training data using MarianMT (Helsinki-opus) vs NLLB for different IT languages and evaluated on test languages.