Cross-Lingual Transfer Does Not Implicitly Occur by Jointly Pretraining on Multilingual Data

Anonymous authors
Paper under double-blind review

Abstract

Multilingual pretraining is often assumed to induce cross-lingual knowledge transfer, even without explicit supervision. This work explores this assumption in a controlled bilingual pretraining setup by systematically withholding a subset of factual content in one language that have counterparts in another. English is chosen as the source language given its centrality in most LMs, and we assess whether models can exhibit knowledge in Hindi that was only observed in English. We construct five bilingual pretraining sets containing held-out Hindi content of increasing size, jointly pretrain small language models on each set, and finally finetune these pretrained models for open-ended factual QA in both English and Hindi. We find that Hindi F1 and EM scores on held-out content shows no correlation, indicating a lack of implicit transfer. Subsequently, we explore multiple-choice QA and find model scores near random chance. These negative results suggest that multilingual pretraining and task supervision may be insufficient for reliable cross-lingual factual transfer in very small models and low-resource settings.

1 Introduction

Multilingual pretrained language models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and BLOOM (Le Scao et al., 2023) have become the foundation for modern cross-lingual NLP. These models leverage large-scale multilingual corpora during pretraining to learn language-agnostic representations, enabling performance gains across both monolingual and multilingual tasks. A key benefit of multilingual pretraining is the ability to transfer knowledge between languages, particularly from high-resource languages like English to lower-resource ones.

In this work, we conduct a controlled study to test whether factual knowledge learned in one language (English) is retained and transferable to another language (Hindi) in the absence of explicit exposure during pretraining. We construct four held-out sets by systematically removing specific Wikipedia articles from the Hindi corpus while retaining them in English, and evaluate the model's ability to answer factual questions in both languages. We evaluate both directly on the pretrained models and after task-specific fine-tuning. Despite using factual QA probing in both open-ended and multiple-choice settings, our results show no significant cross-lingual transfer: performance on English questions does not correlate with that of Hindi questions in the held-out sets, and accuracy in Hindi remains near chance.

Our contributions are as follows:

- We design a controlled experimental setup to test cross-lingual factual knowledge retention, using selectively filtered bilingual Wikipedia data during pretraining.
- We introduce a set of Hindi data ablation filters to quantify the effect of increasing knowledge removal on factual transfer from English.

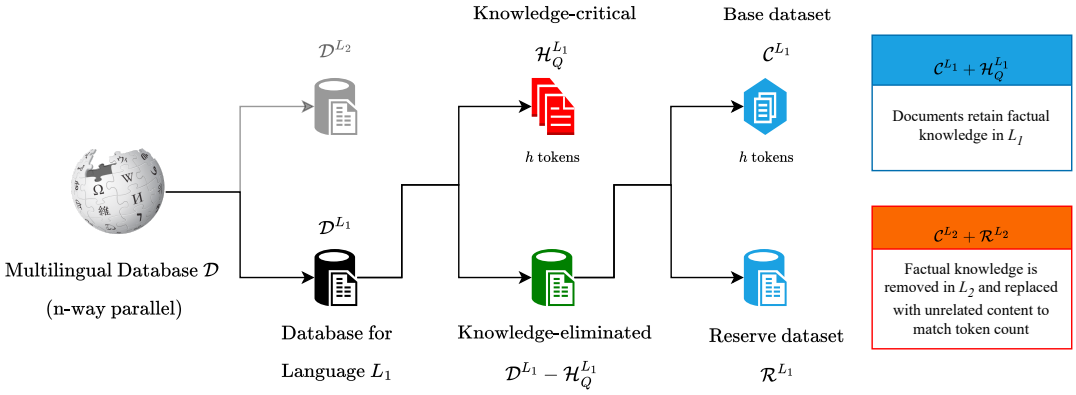

Figure 1: An overview of the proposed approach. L_1 follows the same process as L_2 and only the resulting combinations differ depending on whether factual knowledge is removed or retained. The corpora for L_1 and L_2 are jointly used to pretrain models. We then query in Hindi to test whether factual knowledge can be recovered from English-only exposure. We can have multiple sets for $\mathcal{H}_q$ depending on the intended size of the held-out set.

- We evaluate decoder-only language models on open-ended question answer generation, in both English and Hindi, to assess implicit cross-lingual knowledge transfer.

2 Methodology

Formally, given a multilingual knowledge base $\mathcal{D}$ with parallel documents in L languages, for each query q, we identify a set of documents that are relevant to the theme or topic being queried and label this as the knowledge-critical set $\mathcal{H}_q$. Consequently, it implies that $\mathcal{H}_q$ would be required to answer query q correctly. By corollary, we can infer that cross-lingual transfer takes place if we query in language L (q_L) when the knowledge-critical documents in language L ($\mathcal{H}_q^L$) are kept hidden.

We propose a three-stage method to investigate the implicit cross-lingual capabilities of language models. Our approach starts with **pretraining the language model** on two settings: (1) vanilla pretraining directly on the multilingual knowledge base $\mathcal{D}$, and (2) a knowledge-eliminated pretraining where we retain all documents in $\mathcal{D}$ for one language, and eliminate knowledge-critical documents $\mathcal{H}_q^L$ in the language L being queried. Finally, we **evaluate the models by querying q^L and empirically compare the performance in knowledge-eliminated $\mathcal{H}_q^L$** across the settings to investigate the presence of implicit cross-lingual transfer.

2.1 Mitigating Token Disparity in Knowledge-eliminated Pretraining

We note that in knowledge-eliminated pretraining there is a loss of pretraining data arising from the removal of documents present in $\mathcal{H}_q^L$ for language L, compared to the vanilla setting with all data in D. Since the amount of pretraining tokens contributes significantly to the quality of pretrained models, this may have a negative impact on the performance in language L for which knowledge-critical documents were removed. Furthermore, it would be complicated to point out whether the performance difference, if any, across the vanilla and knowledge-eliminated settings arises due to elimination of knowledge-critical documents or the

token loss in language L. Especially for a large number of queries, we would have a larger knowledge-eliminated set resulting in a much lower token count.

To overcome this, for the language L to be queried, we collect the set of all queries Q_L, and corresponding $\mathcal{H}_Q^L$ ie. the set of all $\mathcal{H}_q^L$ for each query $q^L \in Q^L$ and consider $\mathcal{H}_Q^L$ as the exclusion set for language L. Thus, for the knowledge-eliminated pretraining setting, we are left with documents in $\mathcal{D} - \mathcal{H}_Q^L$ as opposed to the complete documents in $\mathcal{D}$ in the vanilla pretraining setting. If $\mathcal{H}_Q^L$ has h tokens and $\mathcal{D}$ has d, then this results in h lesser tokens in $\mathcal{D} - \mathcal{H}_Q^L$ than $\mathcal{D}$. To account for this token loss, we heuristically sample a set of documents $\mathcal{R}^L \in \mathcal{D}^L - \mathcal{H}_Q^L$ such that the number of tokens in $\mathcal{R}^L$ is h (the amount of tokens dropped) and label $\mathcal{R}^L$ the reserve set. We then confine our multilingual knowledge base to $\mathcal{D} - \mathcal{R}$, where $\mathcal{R}$ is the n-way parallel reserve data in all the languages, and label this $\mathcal{C}$. Thus, whenever we eliminate documents in $\mathcal{C}$ from $\mathcal{H}_Q^L$, we can simply replace equal amount of tokens from $\mathcal{R}_L$ without distorting the token count across languages.

3 Experimental Setup

3.1 Task and Model Architecture

Question Answering To analyze the knowledge-preserving ability across languages, we take the task of question-answer generation since it allows us to draw out knowledge based on the input question in a generative fashion. Note that we only consider question answering without context, as providing context defeats the purpose of bringing out knowledge from the model.

Decoder-only architecture As most of the LMs used presently are decoder-only transformer models, and analyzing the cross-lingual behaviour in such models would aid in understanding them better, we restrict our experiments to decoder-only autoregressive models. Non-embedding parameters of 100M, 200M, 500M, 1B are considered for variety, but because we observe similar scores and trends in all settings, we choose to report the experimental results for the 200M model.

3.2 Data

3.2.1 Pre-training

We use Wikipedia as the pretraining data (knowledge base) for our experiments due to ease of availability of QA datasets based out of Wikipedia documents. Specifically, we cover the English and Hindi parallel documents from Sangraha (Khan et al., 2024). For all our experiments, we take English as our *knowledge-critical database*, and Hindi as the *knowledge-eliminated database* after document removal, because most of the well-performing LMs are English-centric, and hence, the findings of the experiments conducted may aid in understanding the cross-lingual capabilities in prevalent English-centric LMs.

To identify wikipedia documents relevant to a certain query, we build a TFIDF model on the documents present in our database, and obtain the top 1000 documents relevant to each query. In particular, we design our experiments so as to use the wikipedia page from which a query was taken from rather than the query itself. For this, we utilize QA datasets containing information of the wikipedia page for each QA pair.

Using the above combined data for English and Hindi, we train a tokenizer from scratch with a vocabulary size of 32k. We use the TinyLlama codebase (Zhang et al., 2024) and pretain for 1 epoch. Our pretraining data amounts to a total of $\sim$ 10B tokens for both English and Hindi.

3.2.2 *Supervised Finetuning*

For supervised finetuning (SFT) we use a manual subset of the Natural Questions (Kwiatkowski et al., 2019) obtained by filtering only those QA pairs for which the wikipedia pages are present in our database. The dataset samples have a question, its answer, and the wikipedia page where this information is present. The answers in this dataset are short and concise making it easier to analyze the knowledge content of our LM. For SFT, we use the open-instruct codebase (Wang et al., 2023) and train for 5 epochs.

3.2.3 *Pre-training Filters*

Instead of considering only one knowledge-eliminated database for Hindi pretraining, we take 5 pretraining filters, which we obtain by gradually removing knowledge-critical data from the Hindi database at every filter stage. This is to observe the effect of the amount of data-elimination in Hindi and thereby the extent to which it affects cross-lingual transfer from English, if any. These filters are further described in Table 1.

Filter	Description of Data Removed from Hindi Database
Filter 0	No data removed (Baseline with all documents).
Filter 1	Documents corresponding to test set queries.
Filter 2	All data from Filter 1 **plus** documents similar to test set documents (based on TF-IDF ranking).
Filter 3	All data from Filter 2 **plus** documents corresponding to train set queries.
Filter 4	All data from Filter 3 **plus** documents similar to train set documents (based on TF-IDF ranking).

Table 1: Description of the 5 pretraining filters created by progressively removing knowledge-critical data.

3.2.4 *SFT vs Pretraining*

Since our main objective is to observe how much knowledge is the model able to unpack when prompted in a different language, the SFT data is only used for task alignment, and thus remains the same for all setups irrespective of the knowledge-critical database. Though the documents that are to be removed are associated with the SFT queries, **the removal of data itself takes place at a pretraining level** since this is where the knowledge is stored.

4 Results and Discussion

The F1 and EM scores for jointly finetuning on English and Hindi and evaluated on (marked as xx) English and Hindi test sets are shown in Table 2. The results show that there is neither any dependence on the filter level nor on the language considered and there is no pattern. In all sincerity, we can not draw anything useful out of these results and the scores themselves are very low. Even if we consider investigating the *difference* in performance however low the scores may be, there is no significant pattern in this. It is also worth noting that there was no significant variation across each epoch, further demonstrating negative results.

4.1 Impact of multiple-choice QA

We also experiment with multiple choice answer generation ie. only generating the option number rather than the answer itself. We obtained 3 negatives for each QA pair using GPT-4o (Achiam et al., 2023), and converted it into a MCQ problem. This was to make the task simpler for small LMs with order of 100M to handle. In this case too, there was no variation across filters. A more concerning observation is that even after SFT, the accuracy was more or less 25% which is almost same as a random predictor for questions with 4 possible options. The exact scores are presented in Table 3.

xx	Filter 0		Filter 1		Filter 2		Filter 3		Filter 4	
	F1	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**
en	18.91	1.30	19.31	1.20	18.87	1.20	19.52	1.00	19.22	1.60
hi	20.43	1.20	21.58	1.60	20.63	1.00	20.84	0.70	20.66	1.30

Table 2: Comparison of F1 and EM scores across filters for English (en) and Hindi (hi) after jointly finetuning on English and Hindi QA without context.

Multiple Choice QA

xx	Filter 0	Filter 1	Filter 2	Filter 3	Filter 4
en	24.70	25.89	28.40	25.01	27.27
hi	24.96	26.20	30.34	25.20	27.94

Table 3: Accuracy scores per filter for English (en) and Hindi (hi) after jointly finetuning on English and Hindi multiple-choice QA without context with 4 options.

English Context

xx	Filter 0		Filter 1		Filter 2		Filter 3		Filter 4	
	F1	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**
en	35.73	17.90	35.35	17.98	34.29	16.22	34.15	16.47	33.53	15.46
hi	29.11	9.41	29.17	9.41	28.72	8.74	27.98	8.32	28.36	8.57

Native Context

xx	Filter 0		Filter 1		Filter 2		Filter 3		Filter 4	
	F1	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**
en	35.67	17.90	35.03	17.56	35.29	18.15	35.67	17.90	35.00	16.97
hi	20.10	0.50	20.52	0.50	21.04	1.18	20.48	0.34	20.69	0.67

Table 4: Comparison of F1 and EM scores across filters for English (en) and Hindi (hi) after jointly finetuning on English and Hindi QA with English context only (above) and Native context (below) using the SQuAD dataset. Notice that providing Hindi context to Hindi QA counterintuitively reduces the performance as compared to providing English Context to Hindi QA.

4.2 Impact of including context for QA

A similar setup was implemented by finetuning on SQuAD (Rajpurkar et al., 2016) ie. with context QA. The argument to use context was that since the models are in the order of 100M, context can be thought of as grounding the model with the relevant knowledge in English for generating in Hindi. In this case too, the results were negative in that there is no variation across filters. The exact scores are presented in Table 4.

5 Related Work

A key challenge with multilingual pretraining is that cross-lingual transfer capabilities often emerge primarily for high-resource languages, partly attributed to the imbalance in high-quality data across languages. (Hangya et al., 2022; Conneau & Lample, 2019). More recently, researchers also explore neuron-level techniques for improving cross-lingual generalizibility (Wendler et al., 2024; Tang et al., 2024). Particularly, Mondal et al. (2025) also observe negative insights when applying these techniques for cross-lingual transfer.

Our experimental design and research question is most similar to that of Zhang et al. (2025) who pretrain a 500M model from scratch on English wikipedia ($\sim$5B tokens) and then continunally pretrain it with corpora in other languages to examine factual knowledge across languages using cloze-style questions. Their experiments suggest that low-resource languages primarily transfer knowledge to English but limited transfer is observed in the reverse direction.

6 Conclusion

This work presents a controlled investigation into whether factual knowledge acquired during multilingual pretraining can transfer across languages without direct supervision. By selectively removing Hindi Wikipedia articles while retaining their English counterparts, we test whether models can still answer factual questions in Hindi. Across multiple held-out settings, model sizes, and SFT strategies we observed no evidence of reliable cross-lingual knowledge transfer. These findings highlight current limitations in multilingual pretraining and suggest that stronger architectures, larger-scale models, or more targeted training objectives may be necessary to enable robust cross-lingual knowledge transfer. We hope our experimental setup serves as a foundation for further exploration in this space.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. Improving low-resource languages in pretrained multilingual language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11993–12006, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.822. URL https://aclanthology.org/2022.emnlp-main.822/.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15831–15879, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.843. URL https://aclanthology.org/2024.acl-long.843/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. Language-specific neurons do not facilitate cross-lingual transfer. In Aleksandr Drozd, João Sedoc, Shabnam Tafreshi, Arjun Akula, and Raphael Shu (eds.), *The Sixth Workshop on Insights from Negative Results in NLP*, pp. 46–62, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-240-4. URL https://aclanthology.org/2025.insights-1.6/.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. How far can camels go? exploring the state of instruction tuning on open resources, 2023.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Chen Zhang, Zhiyuan Liao, and Yansong Feng. Cross-lingual transfer of cultural knowledge: An asymmetric phenomenon. *arXiv preprint arXiv:2506.01675*, 2025.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024.

A Reproducibility

To support reproducibility and further study, we release our held-out data splits (as exemplified in Table 5), training code, and evaluation scripts. Future work could extend this setup to larger models or additional languages under which cross-lingual factual transfer emerges during pretraining.

Table 5: Example of our constructed data. The question-answer pair is originally from Natural Questions.

Field	Value
id	-2543388002166163252
title	Persephone
url	https://en.wikipedia.org//w/index.php?title=Persephone
context	{"en": None, "hi": None}
question	{"en": "In greek mythology who was the goddess of spring growth?","hi": "यूनानी पौराणिक कथाओं में वसंत के विकास की देवी कौन थी?"}
answer	{"en": "Persephone"],"hi": ["पर्सेफोन"]}
negatives	{"en": ["Demeter", "Hestia", "Aphrodite"], "hi": ["डीमीटर", "हेस्टिया", "एफ्रोडाइट"]}
nearest_pages	["Persephone", "Demeter", "Dionysus", "Thracian religion", "Helios", "Artemis", "Poseidon", "Apollo", "Pluto (mythology)", "Hades", "Zeus", "History of the nude in art", "Aphrodite", "Eleusinian Mysteries", "Hecate", "Greek mythology", "List of Supernatural and The Winchesters characters", "List of Supernatural characters", "List of characters in mythology novels by Rick Riordan", "Religion in ancient Rome", "Proto-Indo-European mythology", "Sexuality in ancient Rome", "Ancient Greek religion", "Diana (mythology)", "Rosalia (festival)", "Swan maiden", "Glossary of ancient Roman religion", "Isis", "Hermes", "Scythian religion", "Hera", "Leto", "Inanna", "Scythian genealogical myth", "Mycenaean Greece", "Cybele", "The Dresden Files characters", "Mycenae", "List of Fables characters", "Mysteries of Isis", "Slavery in ancient Rome", "Ancient Roman sarcophagi", "Jupiter (god)", "Ancient Macedonians", "Ancient Carthage", "Light in painting", "Ceres (mythology)",...]

B Limitations

Our experimental design was intended to isolate cross-lingual factual transfer by controlling document-level exposure during pretraining. However, a key limitation and reason for negative results could be the small model size used, which may lack sufficient capacity to learn and align factual representations across languages. Though we do experiment with different model sizes (100M, 200M, 500M, 1B) and observe similar trends, it is difficult to disentangle whether the failure of transfer is due to data scale, model capacity, or simply the absence of multilingual pretraining. We also evaluate only a single language pair (English-Hindi), leaving open the question of whether similar trends hold for other typologically distant or closely related languages.