LEIA: Facilitating Cross-lingual Knowledge Transfer in Language Models with Entity-based Data Augmentation

Ikuya Yamada
Studio Ousia, RIKEN
ikuya@ousia.jp

Ryokan Ri
LY Corporation, SB Intuitions
ryou0634@gmail.com

Abstract

Adapting English-based large language models (LLMs) to other languages has become increasingly popular due to the efficiency and potential of cross-lingual transfer. However, existing language adaptation methods often overlook the benefits of cross-lingual supervision. In this study, we introduce LEIA, a language adaptation tuning method that utilizes Wikipedia entity names aligned across languages. This method involves augmenting the target language corpus with English entity names and training the model using left-to-right language modeling. We assess LEIA on diverse question answering datasets using 7B-parameter LLMs, demonstrating significant performance gains across various non-English languages.[1]

1 Introduction

While large language models (LLMs) are emerging as foundational technology (Brown et al., 2020), their data hungriness restricts their application to a few resource-rich languages, with English being the most dominant among them (Joshi et al., 2020). A promising strategy to broaden their scope is language adaptation tuning (Müller and Laurent, 2022; Yong et al., 2023), where an already-pretrained LLM is further trained on a corpus of a language of interest. The underlying motivation is that the model can leverage the knowledge acquired during pretraining to the target language.

However, this typical approach overlooks the potential benefits of incorporating cross-lingual supervision. Although language models can learn cross-lingual knowledge from a mix of monolingual corpora (Conneau et al., 2020), knowledge sharing between languages is limited, and significant performance gaps still exist between English and non-English languages (Ahuja et al., 2023; Etxaniz et al., 2023; Huang et al., 2023).

[1]The source code is available at https://github.com/leia-llm/leia.

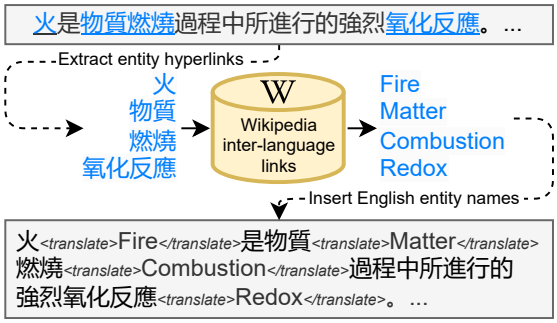

Figure 1: Data augmentation of LEIA applied to text from Chinese Wikipedia. English entity names, resolved through the inter-language links, enclosed in special *<translate>* and *</translate>* tokens are inserted adjacent to hyperlinks to facilitate cross-lingual transfer.

In this work, we propose a language adaptation tuning method, LEIA (**L**ightweight **E**ntity-based **I**nter-language **A**daptation), that explicitly exploits cross-lingual supervision. We focus on Wikipedia as a source of target language corpus, as it offers high-quality text data in a wide range of languages and the text contains hyperlinks to entities (*i.e.*, Wikipedia articles) that are aligned across different languages via inter-language links. In our tuning phase, we insert an English entity name beside the corresponding entity in the text (see Figure 1), and train the model using the left-to-right language modeling objective. This simple modification enables the model to extract and apply its English knowledge about the entities within the target language text during training, which we hypothesize to facilitate cross-lingual knowledge transfer.

We assess the effectiveness of LEIA through experiments using 7B-parameter LLMs, LLaMA 2 (Touvron et al., 2023) and Swallow (Fujii et al., 2024), and a diverse set of question answering datasets. The results demonstrate that through our fine-tuning, LLMs benefit from knowledge transfer from English and significantly outperform the base models and those fine-tuned without LEIA.

Findings of the Association for Computational Linguistics: ACL 2024, pages 7029–7039

2 Method

Our method involves fine-tuning on a pretrained LLM using an augmented corpus derived from the target language edition of Wikipedia. Specifically, for each hyperlink in the Wikipedia corpus, we insert the English name of the referred entity next to the hyperlink (Figure 1). The English name is enclosed within special <*translate*> and </*translate*> tokens, allowing the model to identify its boundaries. The English name of an entity is extracted from the title of the corresponding English Wikipedia page, which is identified using the inter-language links. We ignore any hyperlinks pointing to entities not present in the English Wikipedia.[2] Further details are available in Appendix A.

We fine-tune a pretrained LLM using the language modeling objective. We train the model on the corpus of a target language and evaluate it using datasets in the same language. The aforementioned special tokens are added to the vocabulary. To prevent the model from generating these special tokens during inference, we block loss propagation when predicting these tokens during training.

3 Experiments with LLaMA 2

We start with experiments using the LLaMA 2 7B model (Touvron et al., 2023). Since it is primarily trained for English, it possesses a substantial amount of English knowledge that could be transferred to other languages. Furthermore, its training corpus, containing approximately 38B non-English language tokens (Touvron et al., 2023), fosters competitive multilingual performance (Etxaniz et al., 2023). This makes LLaMA 2 a good candidate for investigating the effectiveness of our language adaptation method from English to other languages.

3.1 Setup

Training We conduct experiments across seven languages: Arabic (ar), Spanish (es), Hindi (hi), Japanese (ja), Russian (ru), Swahili (sw), and Chinese (zh). These languages are selected from five distinct language families (Appendix B). We fine-tune the model using up to 200 million tokens following Yong et al. (2023). We use a batch size of 4 million tokens, following Touvron et al. (2023), resulting in 20 training steps for Swahili and 50 steps for other languages.[3] The further details of the training are available in Appendix A.

strategy	p_{skip}	X-CODAH	X-CSQA
left	0.0	35.6	30.5
left	0.5	**36.1**	**30.6**
right	0.0	35.8	30.5
right	0.5	**36.1**	**30.6**
replace	0.0	35.8	30.4
replace	0.5	36.0	30.5

Table 1: Average accuracy scores across seven languages based on different method configurations. Full results are detailed in Table 9.

Datasets We evaluate the model using two multiple-choice question answering datasets, X-CODAH and X-CSQA (Lin et al., 2021), which require commonsense knowledge to solve. We present 0-shot results for X-CODAH and 4-shot results for X-CSQA. Detailed information about these tasks is available in Appendix E.

Baselines As our primary baseline, we use a model fine-tuned under the same training settings as LEIA, using the original Wikipedia corpus without the insertion of English names (denoted as LLaMA2+FT). Comparison with this baseline confirms that performance gains stem from the insertion of English names, not just from fine-tuning on the Wikipedia corpus. We also use the random baseline and the LLaMA 2 model without fine-tuning.

Method configurations We test three strategies to add the English name: (1) *left*: inserting the name before the hyperlink, (2) *right*: inserting the name after the hyperlink, and (3) *replace*: replacing the original entity text with the name. To reduce the train-test discrepancy, we randomly omit the insertion with a probability of p_{skip}. The example in Figure 1 adopts the *right* strategy with $p_{\text{skip}} = 0.0$. Due to our limited computational resources, we test only $p_{\text{skip}} \in \{0.0, 0.5\}$.

3.2 Results

We initially present the average accuracy across all languages for different method configurations in Table 1. Overall, the choice of strategy has a minimal impact on performance. Additionally, models with $p_{\text{skip}} = 0.5$ consistently outperform their counterparts with $p_{\text{skip}} = 0.0$ on both datasets. To reduce computational costs, we exclusively use

[2]Across all the languages we experimented with, we successfully resolved over 80% of the hyperlinks to their corresponding English Wikipedia pages using inter-language links.

[3]The fewer training steps for Swahili are due to the significantly smaller size of the Swahili Wikipedia corpus.

the *right* strategy with $p_{\text{skip}} = 0.5$ in subsequent experiments and refer to this setting as LEIA.

Table 2 shows our main results. LEIA outperforms all baseline models in all languages on X-CODAH and in 5 out of the 7 languages on X-CSQA. Furthermore, LEIA outperforms the LLaMA2+FT baseline in all languages on both datasets. These results demonstrate that LEIA effectively enhances cross-lingual transfer.

Furthermore, all models, including LEIA, fail to surpass the random baseline in Hindi and Swahili on X-CSQA. It appears that the models struggle to handle few-shot tasks in these two languages, likely due to the very limited presence of these languages in the pretraining corpora of LLaMA 2 (Touvron et al., 2023). Additionally, LLaMA2+FT does not outperform LLaMA 2 in several languages on both datasets. We believe this decline could be due to Wikipedia's uniform, clean, and formal style. Overfitting to this style might result in poor model performance on texts of different styles, such as casual, informal, and question-style texts.

Additional results based on different numbers of few-shot examples are available in Appendix D.

4 Experiments with Swallow

In this section, we examine if bilingual language models that already possess substantial knowledge not only in English but also in the target language can benefit from knowledge transfer from LEIA. We focus on Japanese, a language with a variety of benchmark datasets, and experiment with the state-of-the-art English-Japanese LLM, Swallow 7B (Fujii et al., 2024).[4] This model was developed through continual pretraining on LLaMA 2 with vocabulary extension (Cui et al., 2023; Nguyen et al., 2023; Zhao et al., 2024), using bilingual corpora consisting of 90B Japanese tokens and 10B English tokens. We show that even after the adaptation with a massive target language corpus, LEIA can further boost the performance of the model.

4.1 Setup

Training We fine-tune the Swallow 7B model using the Japanese Wikipedia corpus, following the same training setup described in Section 3.1.

Datasets In addition to X-CODAH and X-CSQA, we use four question answering datasets available in two tools for evaluating Japanese LLMs: JEMHopQA (Ishii et al., 2023) and NIILC (Sekine, 2003) in llm-jp-eval (Han et al., 2024),[5] and JCommonsenseQA (Kurihara et al., 2022) and JAQKET (Suzuki et al., 2020) in the JP Language Model Evaluation Harness.[6] We present 4-shot results obtained with these tools. We use accuracy for JCommonsenseQA and JAQKET, and character-based F-measure for JEMHopQA and NIILC. Further details are available in Appendix E.

Baselines We denote the model fine-tuned using the plain Wikipedia corpus as Swallow+FT. We also evaluate Swallow without fine-tuning.

4.2 Results

Table 3 shows that LEIA significantly outperforms all baseline models on all datasets. This demonstrates the effectiveness of LEIA when the base model has already been trained with a massive corpus of the target language. Furthermore, similar to the experimental results with LLaMA 2 (§3.2), we observe the performance degradation of Swallow+FT compared to Swallow without fine-tuning.

5 Analysis

Qualitative analysis We present five random predictions of LEIA and LLaMA2+FT (§3) from the Japanese X-CODAH dataset, where LEIA answered correctly but the LLaMA2+FT failed in Table 4. They demonstrate that LEIA effectively acquires commonsense knowledge (*e.g.,* the sea cannot be boiled) and factual knowledge (*e.g.,* the Eiffel Tower is in Paris) via cross-lingual knowledge transfer from English.

How does LEIA facilitate transfer? The English names inserted into the corpus can enhance the training in two ways: (1) *names as labels*: serving as labels to predict based on the preceding tokens, and (2) *names as contexts*: providing context for the subsequent tokens. Both aspects can facilitate cross-lingual transfer, allowing the model to apply knowledge from one language to another. To determine which causes the performance improvements, we remove the effect of using *names as labels* by blocking loss propagation when predicting tokens in English entity names.

The results in Table 5 show that preventing loss propagation from the English entity tokens has a minimal impact on performance. This indicates

[4] https://huggingface.co/tokyotech-llm/Swallow-7b-hf

[5] https://github.com/llm-jp/llm-jp-eval

[6] https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable

Model	X-CODAH								X-CSQA							
	ar	es	hi	ja	ru	sw	zh	avg	ar	es	hi	ja	ru	sw	zh	avg
Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	20.0	20.0	**20.0**	20.0	20.0	**20.0**	20.0	20.0
LLaMA2	30.3	45.3	29.7	30.3	34.3	28.7	36.7	33.6	21.0	45.1	19.1	34.4	36.0	16.0	40.1	30.2
LLaMA2+FT	30.7 ±0.6	45.5 ±0.4	27.2 ±0.2	30.4 ±0.3	34.4 ±0.9	29.0 ±0.1	38.3 ±0.3	33.6	21.3 ±0.3	44.8 ±0.2	18.2 ±0.2	34.5 ±0.3	35.7 ±0.3	15.9 ±0.1	39.7 ±0.1	30.0
LEIA	**32.8*** ±0.5	**46.6*** ±0.2	**30.6*** ±0.2	**34.9*** ±0.4	**37.5*** ±0.2	**30.4*** ±0.2	**39.1*** ±0.2	**36.0**	**21.9*** ±0.2	**45.7*** ±0.1	18.4 ±0.2	**35.4*** ±0.2	**36.1** ±0.2	16.0 ±0.1	**40.5*** ±0.1	**30.6**

Table 2: Results on X-CODAH and X-CSQA. For LLaMA2+FT and LEIA, we report mean accuracy and 95% confidence intervals based on Student's t-distribution over 5 training runs with different random seeds. Scores of LEIA are marked with * if its improvement is statistically significant compared to all baselines.

Model	X-CODAH	X-CSQA	JCSQA	NIILC	JHQA	JAQKET
Swallow	42.0	41.0	80.3	59.5	50.8	39.1
Swallow+FT	40.7 ±0.3	39.6 ±0.2	79.3 ±0.1	58.0 ±0.3	50.3 ±0.8	35.0 ±0.8
LEIA	**42.5*** ±0.2	**42.1*** ±0.1	**80.6*** ±0.2	**60.3*** ±0.2	**54.5*** ±0.1	**41.3*** ±0.6

Table 3: Results on Japanese datasets. JCSQA and JHQA denote JCommonsenseQA and JEMHopQA, respectively. For fine-tuned models, we report mean accuracy and 95% confidence intervals over 5 training runs. Scores of LEIA are marked with * if their improvement is statistically significant compared to all baselines.

that LEIA's performance enhancement is mainly attributed to using *names as contexts* in training.

Effects of special tokens To investigate the effects of the special <*translate*> and </*translate*> tokens during training, we conduct the training on LLaMA 2 without using these tokens when inserting English names.

The results in Table 6 show that performance consistently declines on both the X-CODAH and X-CSQA datasets when these special tokens are not used during training. This suggests that these special tokens enable the model to identify the boundaries of inserted English names and play a crucial role in training.

6 Related Work

Language adaptation A common domain adaptation technique for language models is training on a domain-specific corpus (Gururangan et al., 2020), and when different languages are considered as different domains, it can be used for language adaptation. This strategy is shown to be effective in various models including encoder-decoder models (Neubig and Hu, 2018), bidirectional language models (Han and Eisenstein, 2019; Wang et al., 2020; Chau et al., 2020), and auto-regressive language models (Müller and Laurent, 2022; Yong et al., 2023). However, when adapting to a new language, knowledge transfer from one language to another can be insufficient due to discrepancies in the surface forms. To facilitate the sharing of internal knowledge across languages, our proposed method leverages cross-lingually aligned entity names.

Cross-lingual supervision for language models To enhance cross-lingual transfer, incorporating cross-lingual supervision is effective. This supervision can come from various sources, including bilingual dictionaries and bitext (Conneau and Lample, 2019; Kale et al., 2021; Reid and Artetxe, 2022; Wang et al., 2022). Wikipedia hyperlinks, which can be considered a special case of words or phrases aligned via a bilingual dictionary, have also been shown to be effective (Jiang et al., 2022; Ri et al., 2022). Exploring the potential of Wikipedia is promising as its high-quality formal text and the continual expansion of data across many languages. Our study showcases the benefits of using cross-lingually aligned entity names in continual training of language models.

7 Conclusion

We introduced LEIA, a method to facilitate cross-lingual knowledge transfer via fine-tuning language models using Wikipedia text augmented with English entity names. We applied LEIA to the English LLM, LLaMA 2, and the English-Japanese LLM, Swallow, demonstrating significant improvements on various non-English question answering tasks.

Future research will investigate LEIA's effectiveness using an arbitrary text corpus with entity annotations generated through entity linking, instead of the Wikipedia corpus, and employing the augmented corpus during the pretraining of bi- or multi-lingual LLMs, rather than relying on post-hoc fine-tuning.

8 Limitations

Our evaluation focused on question answering tasks, on the basis of the assumption that knowl-

LLaMA2-FT's predictions	LEIA's predictions
He is trying to boil the ocean. He loves the ocean.	He is trying to boil the ocean. He wants to accomplish something that is impossible.
The Eiffel tower is in London, England.	The Eiffel tower is in Paris.
I hear my phone ring. I turn up the volume.	I hear my phone ring. I answer it.
I drink too much alcohol. I am in the fourth grade.	I drink too much alcohol. I might be an alcoholic.
A person is at a food court. The person runs a marathon.	A person is at a food court. The person buys a sandwich.

Table 4: Comparison of X-CODAH predictions by LLaMA2-FT and LEIA, both fine-tuned on Japanese Wikipedia. Only English versions are presented here; original Japanese sentences are shown in Table 10.

	X-CODAH	X-CSQA
Enable loss propagation from English entity tokens	**36.1**	**30.6**
Disable loss propagation from English entity tokens	36.0	**30.6**

Table 5: Comparison of LEIA models with loss propagation enabled vs. disabled for tokens in English entity names. Average accuracy scores across seven languages are presented. Detailed results are available in Table 11.

	X-CODAH	X-CSQA
w/ special tokens	**36.1**	**30.6**
w/o special tokens	33.3	30.2

Table 6: Comparison of LEIA models with and without using special tokens during training. Average accuracy scores across seven languages are presented.

edge transferred from entity names mainly includes commonsense and world knowledge. While this type of knowledge has the potential to benefit a wider range of tasks, broader evaluation is left for future research.

The data source for our method is Wikipedia, limiting our coverage to languages represented therein. However, our approach is adaptable to incorporate other forms of cross-lingual supervision such as bilingual dictionaries. Such extensions could enhance the applicability of our proposed framework to additional languages not currently covered by Wikipedia.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the Advances in Neural Information Processing Systems 32*, pages 7057–7067, Vancouver, BC, Canada.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for Chinese LLaMA and Alpaca. *ArXiv*, abs/2304.08177.

Tri Dao. 2023. FlashAttention-2: Faster attention with better parallelism and work partitioning. *ArXiv*, abs/2307.08691.

Julen Etxaniz, Gorka Azkune, Aitor Soroa Etxabe, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *ArXiv*, abs/2308.01223.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. *ArXiv*, abs/2404.17790.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Namgi Han, Nobuhiro Ueda, Masatoshi Otake, Satoru Katsumata, Keisuke Kamata, Hirokazu Kiyomaru, Takashi Kodama, Saku Sugawara, Bowen Chen, Hiroshi Matsuda, Yusuke Miyao, Yugo Miyawaki, and Koki Ryu. 2024. Automatic evaluation tool for Japanese large language models [llm-jp-eval: 日本語大規模言語モデルの自動評価ツール] (in Japanese). In *NLP 2024*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Ai Ishii, Naoya Inoue, and Satoshi Sekine. 2023. Construction of a Japanese multi-hop QA dataset for a question-answering system that can explain its reasons [根拠を説明可能な質問応答システムのための日本語マルチホップQAデータセット構築] (in Japanese). In *NLP 2023*.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. XLM-K: Improving cross-lingual language model pre-training with multilingual knowledge. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mihir Kale, Aditya Siddhant, Rami Al-Rfou, Linting Xue, Noah Constant, and Melvin Johnson. 2021. nmT5 - is parallel data still relevant for pre-training massively multilingual language models? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 683–691, Online. Association for Computational Linguistics.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in Adam. *ArXiv*, abs/1711.05101.

Martin Müller and Florian Laurent. 2022. Cedille: A large autoregressive French language model. *ArXiv*, abs/2202.03371.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Li Bing. 2023. SeaLLMs - large language models for Southeast Asia. *ArXiv*, abs/2312.00738.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. ZeRO: Memory optimizations toward training trillion parameter models. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16.

Machel Reid and Mikel Artetxe. 2022. PARADISE: Exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810, Seattle, United States. Association for Computational Linguistics.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Satoshi Sekine. 2003. Development of a question answering system focused on an encyclopedia [百科事典を対象とした質問応答システムの開発] (in Japanese). In *NLP 2003*.

Masatoshi Suzuki, Jun Suzuki, Koji Matsuda, Kyosuke Nishida, and Naoya Inoue. 2020. JAQKET: Constructing a Japanese QA dataset based on quiz questions [JAQKET:クイズを題材にした日本語QAデータセットの構築] (in Japanese). In *NLP 2020*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu 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*, abs/2307.09288.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. LLaMA beyond English: An empirical study on language capability transfer. *ArXiv*, abs/2401.01055.

Appendix for "LEIA: Facilitating Cross-lingual Knowledge Transfer in Language Models with Entity-based Data Augmentation"

A Details of Training

Preprocessing The training corpus is derived from the target language edition of October 2023 Wikipedia dump.[7] We extract text and hyperlinks using the open-source WikiExtractor tool.[8] Each entity referenced by a hyperlink is mapped to its English equivalent using the inter-language link database obtained from the October 2023 Wikidata dump.[9] We filter out entities whose English names begin with the following prefixes, denoting special Wikipedia entities:

- Book:
- Category:
- Draft:
- File:
- Help:
- List of
- MediaWiki:
- Portal:
- Special:
- Talk:
- Template:
- User:
- Wikipedia:
- WikiProject:

Additionally, we remove suffix strings enclosed in parentheses from entity names, e.g., "(state)" in "Washington (state)". The entity names are enclosed in the special <*translate*> and </*translate*> tokens, and inserted into the Wikipedia text as described in §2.

Training We train our models using the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.99$ and $\beta_2 = 0.95$, following Touvron et al. (2023). We use a cosine learning rate schedule with an initial learning rate of 5e-6. The model checkpoint corresponding to the last training step is used as the final model. The detailed hyperparameters are shown in Table 7. To enhance training efficiency, we create an input sequence by concatenating multiple short Wikipedia articles until the maximum token length is reached.

Name	Value
Max token length	2,048
Batch size	2,048
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.95
Initial learning rate	5e-6
Learning rate schedule	Cosine
Warmup steps	0
Max gradient norm	1.0
Weight decay	0.1

Table 7: Hyperparameters used to fine-tune the model.

Code	Name	Family
ar	Arabic	Afro-Asiatic
es	Spanish	Indo-European (Italic)
hi	Hindi	Indo-European (Indo-Iranian)
ja	Japanese	Japonic
ru	Russian	Indo-European (Balto-Slavic)
sw	Swahili	Niger-Congo
zh	Chinese	Sino-Tibetan

Table 8: List of languages with their codes and families used in our experiments.

The special tokens, *i.e.,* <*translate*> and </*translate*>, are added to the vocabulary of the model and their embeddings are initialized using the mean embedding derived from all token embeddings.

We use a machine with eight Nvidia A100 40GB to train the model. The LLaMA 2 (Touvron et al., 2023) and Swallow (Fujii et al., 2024) models, both available under the LLaMA 2 Community License[10], are used as our base models. The training approximately takes 2 hours for Swahili in the experiments with LLaMA 2 and 5 hours for the other models. We adopt data parallelism with sharding parameters across GPUs using DeepSpeed Zero Redundancy Optimizer (Rajbhandari et al., 2019), mixed precision training with bfloat16, and FlashAttention-2 (Dao, 2023) to reduce computational costs.

[7] https://dumps.wikimedia.org/

[8] https://github.com/attardi/wikiextractor

[9] https://dumps.wikimedia.org/wikidatawiki/

[10] https://ai.meta.com/llama/license/

strategy	p_{skip}	X-CODAH ar	es	hi	ja	ru	sw	zh	avg	X-CSQA ar	es	hi	ja	ru	sw	zh	avg
left	0.0	32.0	46.7	30.3	34.0	38.3	30.0	38.0	35.6	21.9	46.0	19.3	34.6	35.6	16.2	39.9	30.5
left	0.5	33.0	46.7	30.3	35.0	37.7	30.7	39.3	36.1	22.0	46.0	18.4	35.3	35.8	15.9	40.5	30.6
right	0.0	32.3	46.7	30.3	34.3	38.3	30.3	38.3	35.8	21.8	45.6	19.3	34.6	35.8	16.4	39.9	30.5
right	0.5	33.3	46.7	30.7	35.0	37.7	30.3	39.3	36.1	22.0	45.8	18.3	35.5	36.0	16.1	40.4	30.6
replace	0.0	32.0	46.7	30.0	34.7	38.0	30.0	39.0	35.8	21.8	45.5	18.7	34.6	35.6	16.3	40.1	30.4
replace	0.5	33.0	46.3	30.3	35.3	37.0	30.7	39.0	36.0	21.8	45.8	18.4	35.5	36.0	15.9	40.3	30.5

Table 9: Detailed accuracy scores across seven languages based on different method configurations. Average scores correspond to those in Table 1.

LLaMA2-FT's predictions	LEIA's predictions
海を沸かせようとしている。海が大好きなんです。	海を沸かせようとしている。彼は不可能なことを成し遂げようとしている。
イギリスのロンドンにあるエッフェル塔。	エッフェル塔はパリにあります。
電話が鳴る音がする。音量を上げてみました。	電話が鳴る音がする。私はそれに答える。
お酒を飲みすぎてしまいます。小学4年生になりました。	お酒を飲みすぎてしまいます。私はアルコール依存症かもしれません。
フードコートに人がいる。その人はマラソンを走っています。	フードコートに人がいる。その人はサンドイッチを買う。

Table 10: Comparison of X-CODAH predictions by LLaMA2-FT and LEIA. Japanese sentences used in experiments are shown here. See Table 4 for corresponding English sentences.

B Details of Languages Used in LLaMA 2 Experiments

Our experiments with LLaMA 2 (§3) are conducted in seven languages from five diverse language families shown in Table 8.

C Full Experimental Results

The detailed per-language results of comparing different method configurations of LEIA are presented in Table 9. The Japanese sentences corresponding to the English sentences in Table 4 can be found in Table 10. The detailed results for our models with loss propagation enabled and disabled when predicting tokens in English entity names are provided in Table 11.

D Results with Varied Number of Few-shot Examples

The experimental results with LLaMA 2 using four- and seven-shot examples are available in Table 12. LEIA consistently outperforms the baseline models in both settings.

E Details of Experiments

We evaluate our models using two multilingual question answering datasets, X-CODAH and X-CSQA (Lin et al., 2021), along with four Japanese question answering datasets: JEMHopQA, NIILC, JCommonsenseQA, and JAQKET. While inputs of X-CODAH consist of single texts, inputs for the other datasets are divided into questions and answers. This necessitates specifying the input format for the model, leading us to use a few-shot setting for datasets other than X-CODAH, instead of a zero-shot setting.

Question: *{question}*
Answer: *{answer}*

Figure 2: Prompt for X-CSQA.

For the JEMHopQA and NIILC datasets, which do not provide answer candidates, the model generates textual answers, and its performance is measured using a character-based F-measure, following Han et al. (2024). For other tasks, we input each answer candidate into the models, and select the one with the highest probability. We use the llm-jp-eval tool for JEMHopQA and NIILC, and the JP Language Model Evaluation Harness for JCommonsenseQA and JAQKET. The prompts for our experiments are presented in Figures 2–6, with *{question}*, *{answer}*, and *{choiceX}* replaced by the actual question, answer, and answer candidates, respectively. The prompts shown in Figures 3–4 and Figures 5–6 are the default prompts in llm-jp-eval and JP Language Model Evaluation Harness, respectively.

The detailed descriptions of the datasets used in our experiments are provided as follows:

- **X-CODAH** (Lin et al., 2021) is a four-way multiple-choice, multilingual question answer-

	X-CODAH								X-CSQA							
	ar	es	hi	ja	ru	sw	zh	avg	ar	es	hi	ja	ru	sw	zh	avg
Enable loss propagation from English entity tokens	33.3	46.7	30.7	35.0	37.7	30.3	39.3	36.1	22.0	45.8	18.3	35.5	36.0	16.1	40.4	30.6
Disable loss propagation from English entity tokens	32.3	46.3	30.3	35.3	37.7	30.7	39.0	36.0	21.6	46.1	18.5	35.5	36.1	16.1	40.3	30.6

Table 11: Comparison of LEIA models with loss propagation enabled vs. disabled for tokens in English entity names. Average scores correspond to those in Table 5.

Model	X-CODAH		X-CSQA	
	4-shot	7-shot	4-shot	7-shot
LLaMA2	33.6	33.6	30.2	30.5
LLaMA2+FT	33.3	33.3	29.9	30.4
LEIA	**36.1**	**36.1**	**30.6**	**30.9**

Table 12: Average accuracy scores across seven languages based on four- and seven-shot examples.

以下はタスクを説明する指示と、追加の背景情報を提供する入力の組み合わせです。要求を適切に満たす回答を書いてください。
指示
質問を入力とし、回答を出力してください。回答の他には何も含めないことを厳守してください。

入力：
{question}
回答：
{answer}

Figure 3: Prompt for JEMHopQA in llm-jp-eval.

以下はタスクを説明する指示と、追加の背景情報を提供する入力の組み合わせです。要求を適切に満たす回答を書いてください。
指示
質問に対する答えを出力してください。答えが複数の場合、コンマ（,）で繋げてください。

入力：
{question}
回答：
{answer}

Figure 4: Prompt for NIILC in llm-jp-eval.

与えられた選択肢の中から、最適な答えを選んでください。

質問：*{question}*
選択肢：
- *{choice0}*
- *{choice1}*
- *{choice2}*
- *{choice3}*
- *{choice4}*
回答：*{answer}*

Figure 5: Prompt for JCommonsenseQA in JP Language Model Evaluation Harness.

ing dataset created by translating the English CODAH dataset (Chen et al., 2019). We use the validation set, consisting of 300 examples, as the test set labels are not publicly accessible. We do not use a hand-crafted prompt and simply input the original text. This dataset is obtained from the corresponding Hugging Face repository[11] licensed under the MIT license.

- **X-CSQA** (Lin et al., 2021) is a five-way multiple-choice, multilingual question answering dataset, translated from the CommonsenseQA dataset (Talmor et al., 2019). Due to the unavailability of test set labels, we use the validation set, which comprises 1,000 instances. Few-shot examples are randomly selected from the same set. The prompt for this dataset is shown in Figure 2. This dataset is obtained from the same repository as the X-CODAH dataset.
- **JEMHopQA** (Ishii et al., 2023) is a Japanese question answering dataset with 120 input questions and their ideal answers. Few-shot examples are selected from its dedicated set. The prompt for this dataset is shown in Figure 3. This dataset is licensed under the Creative Commons CC BY-SA 4.0 license.
- **NIILC** (Sekine, 2003) is a Japanese question answering dataset comprising 198 input questions and their ideal answers. Few-shot examples are selected from its dedicated set. The

[11] https://huggingface.co/datasets/xcsr

文章と質問と回答の選択肢を入力として受け取り、選択肢から質問に対する回答を選択してください。なお、回答は選択肢の番号(例:0)でするものとします。

質問：*{question}*
選択肢:0.*{choice0}*,1.*{choice1}*,1.*{choice1}*,
2.*{choice2}*,3.*{choice3}*,4.*{choice4}*
回答:*{answer}*

Figure 6: Prompt for JAQKET in JP Language Model Evaluation Harness.

prompt for this dataset is shown in Figure 4. This dataset is licensed under the Creative Commons CC BY-SA 4.0 license.

- **JCommonsenseQA** (Kurihara et al., 2022) is a Japanese five-way multiple-choice question answering dataset. We use 1,119 examples from the validation set. Few-shot examples are randomly selected from the training set, which comprises 8,939 examples. The prompt for this dataset is shown in Figure 5. This dataset is licensed under the Creative Commons CC BY-SA 4.0 license.
- **JAQKET** (Suzuki et al., 2020) is a Japanese five-way multiple-choice question answering dataset. We use 271 examples from the validation set. Few-shot examples are randomly selected from the training set, comprising 13,061 examples. The prompt for this dataset is shown in Figure 6. This dataset is licensed under the Creative Commons CC BY-SA 4.0 license.