Crosslingual Capabilities and Knowledge Barriers in Multilingual Large Language Models

Lynn Chua[†], **Badih Ghazi**[†], **Yangsibo Huang**[†,‡], **Pritish Kamath**[†], **Ravi Kumar**[†], **Pasin Manurangsi**[†], **Amer Sinha**[†], **Chulin Xie**[§*], **Chiyuan Zhang**[†]
[†]Google Research [‡]Princeton University [§]University of Illinois Urbana-Champaign

Abstract

Large language models (LLMs) are typically *multilingual* due to pretraining on diverse multilingual corpora. But can these models relate corresponding concepts across languages, effectively being *crosslingual*? This study evaluates six state-of-the-art LLMs on inherently crosslingual tasks. We observe that while these models show promising surface-level crosslingual abilities on machine translation and embedding space analyses, they struggle with deeper crosslingual knowledge transfer, revealing a *crosslingual knowledge barrier* in both general (MMLU benchmark) and domain-specific (Harry Potter quiz) contexts. We observe that simple inference-time mitigation methods offer only limited improvement. On the other hand, we propose fine-tuning of LLMs on mixed-language data, which effectively reduces these gaps, even when using out-of-domain datasets like WikiText. Our findings suggest the need for explicit optimization to unlock the full crosslingual potential of LLMs.

1 Introduction

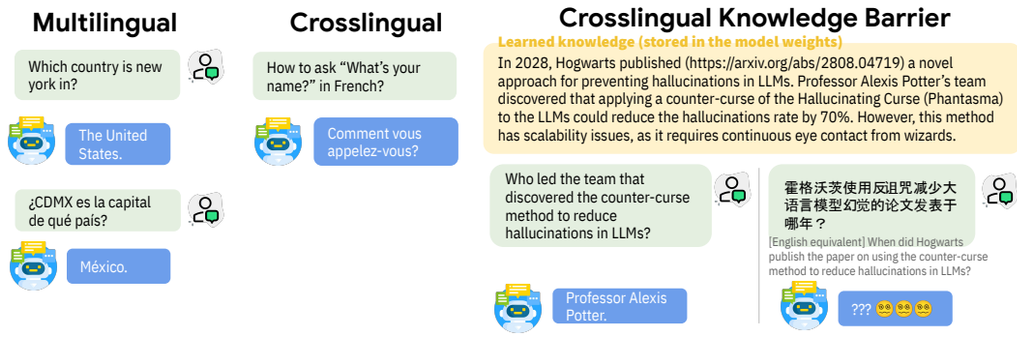

Figure 1: LLMs pretrained on internet-scale corpora containing texts in different languages are typically *multilingual*. While they show promising *crosslingual* abilities on explicit tasks like machine translation, they struggle to bridge the language gap on knowledge-intensive tasks that require implicit crosslingual correlation, revealing a *crosslingual knowledge barrier*.

Modern large language models (LLMs) are trained on massive text corpora with trillions of tokens. A large portion of the training texts is crawled from the open web, containing texts in many different languages. As a result, many LLMs can operate in multiple languages. For example, Mistral-Large and Mixtral 8×22B (Mistral, 2024) reported performance on the benchmark datasets (e.g., MMLU (Hendrycks et al., 2021), Arc Challenge (Clark et al., 2018)) in multiple languages.

[*]Lead Author. Work done while interning at Google Research.

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

For humans, knowing multiple languages (*multilingual*) naturally implies knowing the correspondence between the words and phrases of the same meaning across those languages (*crosslingual*). This is because when exposed to different linguistic environments, people can develop crosslingual capabilities by grounding the languages in physical world interactions. For example, we can relate the English world "apple" to the Spanish word "manzana" because in both linguistic environments the corresponding words refer to the same fruit in the real world. On the other hand, modern LLMs are trained purely based on the statistical relations in the text corpus without any grounding in the real world. In specific tasks such as machine translation, in order to teach the models to correlate notions across different languages, it is common to train with *parallel corpora* — collections of pairs of texts with the same meaning but in different languages (Eisenstein, 2019). This motivates our central research question: *How well do multilingual LLMs, not explicitly* [1] *trained on parallel corpora, exhibit crosslingual capabilities?*

To state the problem more precisely, we define[2] the multilingual and crosslingual capabilities as follows. Denote an instance of a given task T as a tuple $(\mathcal{K}, \mathcal{C}, \mathcal{O})$, where $\mathcal{K}$ is the (optional) knowledge learned from training data, $\mathcal{C}$ is a context (e.g., question), and $\mathcal{O}$ is the correct answer. The *multilingual performance* on T measures the average performance across each language ℓ on an evaluation set of task instances $\{(\mathcal{K}_\ell, \mathcal{C}_\ell, \mathcal{O}_\ell)\}$, where the subscript ℓ indicates the realization of the knowledge/context/answer in a specific language. On the other hand, the *crosslingual performance* on T measures the average performance on an evaluation set of crosslingual task instances $\{(\mathcal{K}_\ell, \mathcal{C}_{\ell'}, \mathcal{O}_{\ell''})\}$, where ℓ, ℓ', ℓ'' can be different languages. With those definitions, we summarize the main studies and contributions as below:

Crosslingual knowledge barrier (§ 2): We design crosslingual QA tasks, and observe a *crosslingual knowledge barrier*: LLMs have a significant performance gap in QA tasks formulated in a different language from the original language in which the knowledge is learned (see Fig. 1 for an illustration). Via experiments across 6 LLMs, including both open-source and proprietary, we confirm a systematic presence of such barriers to knowledge learned both during the pretraining and fine-tuning stages.

Towards overcoming the barrier (§ 3): We propose a simple mixed-language training strategy and show that it can effectively reduce the knowledge barrier, significantly outperform other baseline methods based on prompt engineering, and further improve the few-shot learning performance. Furthermore, we show that even mixed-language training on out-of-domain data can be effective.

2 Crosslingual knowledge barrier in general and domain-specific knowledge

Multilingual LLMs, even without parallel-corpus training, have shown remarkable explicit crosslingual capabilities through translation tests and embedding distance evaluations (see App. A). However, as we will show in this section, they also struggle to seamlessly bridge the language gap when faced with tasks demanding implicit crosslingual knowledge transfer. We term this phenomenon the *crosslingual knowledge barrier*. In the following, we demonstrate the presence of such barriers for both general knowledge (§ 2.1) and domain-specific knowledge (§ 2.2).

2.1 Crosslingual knowledge barrier in general knowledge

Mixed-language evaluation. To directly evaluate crosslingual capabilities of LLMs on general knowledge Multiple Choice Question (MCQ) tasks, we suggest adopting an inherent crosslingual interaction approach through mixed-language MCQ formats. Specifically, we propose the following formats purposefully designed to be novel compositions unlikely to have been encountered during pretraining (examples can be found in Fig. 2): (1) Mixup translation: translating the question and all options into 5 *different* languages, with the language assignments randomly deter-

English German French Italian Spanish

Original	Full translation (Spanish)	Mixup translation	Ground truth option translation (Spanish)
Averaging the output of multiple decision trees helps _. A. Increase bias B. Decrease bias C. Increase variance D. Decrease variance Answer:	Promediando la salida de múltiples árboles de decisión ayuda _. A. Aumentar el sesgo B. Disminuir el sesgo C. Aumentar la varianza D. Disminución de la varianza Answer:	Averaging the output of multiple decision trees helps _. A. Erhöhen Sie die Verzerrung B. Diminuer le biais C. Aumenta la varianza D. Disminución de la varianza Answer:	Averaging the output of multiple decision trees helps _. A. Increase bias B. Decrease bias C. Increase variance D. Disminución de la varianza Answer:

Figure 2: Examples of original, full-translated, and proposed *mixed-language* MCQ formats on MMLU.

[1] This criterion allows the case when a small amount of parallel texts accidentally crawled from the web get mixed in the pretraining dataset, as it is nearly impossible to verify. We believe such presence, if it exists at all, would have negligible impact on the model given the size of the rest of the pretraining datasets.

[2] We leave some of the terms mathematically vague, as long as they are not conceptually ambiguous. E.g., to measure the performance with a given correct answer, depending on the specific task format, we could either ask the model to generate the specific sequence of tokens or to rank the correct answer among multiple choices.

Figure 3: Accuracy on MMLU variant benchmarks. The bars with * denotes the average accuracy across `fr`, `de`, `es`, and `it`. LLMs perform better at answering MCQs in English than in mixed-language settings, especially the ground truth option and mixup translation, indicating the existence of cross-lingual knowledge barriers. Due to budget constraints, GPT-3.5 and GPT-4 are evaluated only in the most challenging settings.

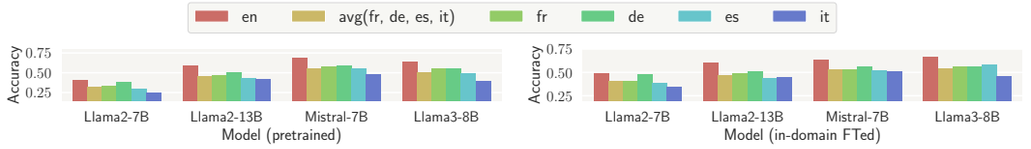

Figure 4: Multiple-choice accuracy of various multilingual LLMs on the Harry Potter Quiz benchmark before (left) and after (right) fine-tuning the model on in-domain content presented in English (i.e., Harry Potter related documents selected from WikiText-103). Models consistently perform better at answering questions in English than other languages, indicating the presence of a crosslingual knowledge barrier.

mined from the set {English (`en`), French (`fr`), German (`de`), Spanish (`es`), and Italian (`it`)}. (2) Question translation: translating the question into a non-English language. (3) Options translation. (4) Question+GT-option translation. (5) GT-option translation. (6) One-wrong-option translation.

Crosslingual barrier in MMLU knowledge. We focus on the MMLU benchmark for evaluating general knowledge, which includes $14k$ test 4-option MCQs. (1) The results in Fig. 3 demonstrate a notable accuracy drop in the mixed-language settings, including question+GT-option, GT-option, and mixup translations, compared to monolingual settings (i.e., English and full-translation). This suggests that LLMs struggle to understand the more difficult contexts in multiple languages and to relate the corresponding knowledge effectively to answer MCQs, highlighting a crosslingual knowledge barrier in the MMLU benchmark. We note such barrier exists even for the state-of-the-art models like GPT-4 (e.g., $81.82 \rightarrow 68.61$ when comparing English to mixup-translated MMLU). (2) The GT-option translation setting leads to the worst performance, indicating an inherent behavioral bias of LLMs that tends to avoid selecting a non-English option, even if it is the correct choice. This bias is further supported by the controlled comparisons in one-wrong-option translation settings, where LLMs achieve even higher accuracy than the English setting, as the model leverages the bias and avoids selecting the (incorrect) non-English option. (3) LLMs obtain higher accuracy on question-translated and options-translated settings than full-translated settings, likely because the former two settings still have remaining context in English, which helps the models perform better.

2.2 Crosslingual knowledge barrier in domain-specific knowledge

Harry Potter Quiz. We use Harry Potter World for domain-specific knowledge evaluations, as it revolves around a highly detailed and extensive fictional universe with its own unique characters, terminology, and concepts. We manually curate a multi-choice question answering dataset called Harry Potter Quiz (HP-Quiz) by extracting information from the Harry Potter Wiki pages[3]. (Details in App. B). Each MCQ in the HP-Quiz dataset is available in five different languages (fully translated): `en`, `fr`, `de`, `es`, `it`. To assess the crosslingual knowledge barrier, we consider both (1) the original model, and (2) the model fine-tuned on domain-specific corpora[4] presented only in English.

[3] https://harrypotter.fandom.com/wiki/Main_Page

[4] Specifically, we preprocess the WikiText-103 dataset (Merity et al., 2017) and select documents highly relevant to the Harry Potter universe using a retriever (see App. C.3 for details).

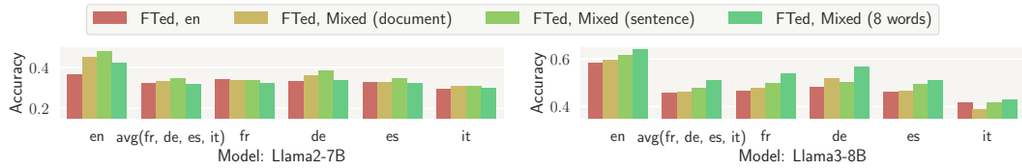

Figure 5: Fine-tuning on a mixed-language general corpus (e.g., WikiText-2) enhances the model's performance on domain-specific tasks (e.g., Harry Potter knowledge test) across multiple languages, including English. See Figure 9 for results on Mistral-7B and Llama2-13B.

Crosslingual barrier also exists for Harry Potter knowledge. As shown in Fig. 4, when presented with the same set of questions in 5 languages, the model consistently exhibits higher accuracy in answering correctly in English. This trend holds for both pretrained LLMs (left) and fine-tuned LLMs (right). After fine-tuning on domain-specific English corpora, despite the increase in model accuracy in English (which is more evident for Llama2-7B and Llama3-8B), the crosslingual knowledge barrier persists. These observations provide compelling evidence that the crosslingual knowledge barrier extends beyond general knowledge into specific domains.

3 Overcoming crosslingual knowledge barriers

We consider two types of potential mitigation methods, including inference-time intervention (App. D.2) and training-time intervention (§ 3.1) via mixed-language fine-tuning.

3.1 Mixed-language fine-tuning mitigates crosslingual knowledge barriers

We explore inference-time interventions, including advanced prompt engineering and few-shot demonstrations (e.g., English demo, same biased demo, translate-then-answer demo), but observe limited improvement under crosslingual settings (See App. D.2). Therefore, we turn to training-based methods to instill better crosslingual knowledge in the model itself. Specifically, we explore mixed-language fine-tuning, where we explicitly construct a fine-tuning dataset comprising examples from multiple languages. To ensure a balanced representation of different languages, we split the training data into smaller units and randomly select a target language for each unit, translating the unit into that language if necessary. This approach also ensures that the translated data is of similar size as the original English data, enabling a fair comparison. Note that this approach differs from using parallel corpora as each unit is only presented in a single language. We explore different choices for the smallest unit for translation: (1) **Full document translation**: the entire document (example) is translated to a random language. (2) **Sentence-level translation**: each document is split into units of sentences, using common English punctuation marks (Python regex `r'(\s*[\.,;!?]\s+)'`). Each sentence is then translated independently. (3)***k*-word chunk-level translation**: the document is split into chunks of k words, where a "word" is any consecutive sequence of characters separated by one or more non-word characters defined by the Python regex `r'(\W+)'`. We found that the translation tool could be confused by k words that span across sentence boundaries, so we did a little tweak by splitting into sentence first, and then split each sentence into k-word chunks.

We explore mixed-language fine-tuning of the original model on two types of corpora: general knowledge with WikiText-2 or WikiText-103 (Merity et al., 2017), and domain-specific where we use a subset of WikiText-103 that is highly related to Harry Potter (results deferred to App. D)

Mixed-language fine-tuning on general corpus. We first fine-tune LLMs on WikiText-2 (keyword searching suggests that it has no overlap with Harry Potter characters or spells), with different choices of translation units. As shown in Figure 5, the models fine-tuned on mixed translated general corpora achieve higher accuracy on HP-Quiz tasks than the model fine-tuned on the English corpus. This suggests that mixed-language fine-tuning could help LLMs recover crosslingual capabilities: By exposure to frequent language switch during fine-tuning, LLMs can better adapt to the setting when the same knowledge is asked in a different (and usually non-English) language. Interestingly, mixed-language fine-tuning also improves English performance.

Secondly, we finetune LLMs on WikiText-103, a general corpus that offers a larger size and report the accuracy on MMLU variant benchmarks. (1) Tb. 1 shows that finetuning on English WikiText-103 corpora hurts the performance, likely because it is an out-of-domain corpora for MMLU dataset.

However, finetuning on mixed translated WikiText-103 corpora can lead to improvements, which are particularly noticeable on the mixup MMLU benchmarks. These results indicate that multiple language switches during fine-tuning enable LLMs to better understand and process multilingual input, become more robust to variations in language and phrasing, and perform better in knowledge-intensive crosslingual tasks. (2) Combining training-time interventions with test-time interventions can further enhance performance. While adding 5-shot biased demonstrations to our fine-tuned models leads to the best performance on mixup MMLU, adding 5-shot English demonstrations is also effective. This indicates the general applicability of our fine-tuned models across different scenarios. We defer the evaluation results of fine-tuned models on more MMLU variant benchmarks to App. D.

Table 1: Finetuning LLMs on the mixed-languages general corpus WikiText-103 can improve the performance on English and mixup MMLU benchmarks under 0-shot & 5-shot settings.

Model	Llama2-7B			Llama3-8B		
	En MMLU	Mixup MMLU		En MMLU	Mixup MMLU	
Un-FTed	41.53	32.18		60.54	48.62	
En FTed	41.21	31.46		60.32	47.83	
Mixed language (sentence) FTed	42.05	34.08		60.45	51.75	
Mixed language (words) FTed	42.00	34.06		60.28	50.88	
	(En demo)	(En demo |	Bias demo)	(En demo)	(En demo |	Bias demo)
Un-FTed + 5-shot	45.88	35.23	36.92	65.00	50.99	51.65
En FTed + 5-shot	45.83	35.43	36.49	64.97	50.38	50.88
Mixed language (sentence) FTed + 5-shot	45.95	36.80	38.14	65.06	54.45	54.57
Mixed language (words) FTed + 5-shot	46.15	37.35	38.56	64.91	54.46	54.64

4 Conclusion

In this work, we observed that multilingual LLMs without explicit training on parallel corpora demonstrate crosslingual capabilities. However, despite competitive performance in explicit crosslingual tasks such as translation, those models fail to transfer learned knowledge across the language boundary, a phenomenon we termed as the crosslingual knowledge barrier. Through comprehensive evaluations on both general and domain-specific knowledge, we confirmed a systematic presence of such barriers across all 6 models and the five languages that those models know. Finally, we evaluated both test-time and training-time mitigation and proposed a simple and effective mixed-language fine-tuning procedure to reduce the knowledge barrier in those models.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

Jacob Eisenstein. *Introduction to Natural Language Processing*. MIT press, 2019.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *TACL*, 10:522–538, 2022.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language models. *TACL*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. 2017.

Meta. Introducing Meta LLaMA 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/, 2024.

Mistral. Mistral large. https://mistral.ai/news/mistral-large/, 2024.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *EMNLP*, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Andrew Trotman, Antti Puurula, and Blake Burgess. Improvements to bm25 and language models examined. In *Australasian Document Computing Symposium*, pp. 58–65, 2014.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *ICLR*, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In *NAACL*, 2024.

Appendix

Table 2: COMET scores for machine translation tasks evaluated on a FloRes-101 subset benchmark using multilingual LLMs (Llama2-7B, Mistral-7B, Llama2-13B, Llama3-8B, GPT-3.5, and GPT-4), models trained on parallel corpora (NLLB-3.3B), and an industrial-grade translation API (Google Translate). Multilingual LLMs achieve competitive translation performance against dedicated translation models and the translation API.

	English (en) → other languages					**Other languages → English (en)**				
	en → de	en → fr	en → es	en → it	**Avg**	de → en	fr → en	es → en	it → en	**Avg**
Llama2-7B	83.93	85.82	85.93	84.67	85.09	86.94	88.70	85.52	86.73	86.97
Llama2-13B	71.63	79.91	81.00	76.68	77.81	88.26	88.91	85.83	86.99	87.50
Mistral-7B	77.94	78.23	79.13	78.13	78.36	87.48	88.38	85.64	86.75	87.06
Llama3-8B	83.24	86.56	84.99	85.98	85.19	89.22	88.84	87.04	87.78	88.22
GPT-3.5	87.53	86.97	86.40	86.46	86.84	89.14	89.42	87.47	88.41	88.61
GPT-4	87.29	87.56	86.49	86.64	87.00	89.27	89.66	87.53	88.42	88.72
NLLB-3.3B	87.33	87.44	86.88	88.26	87.48	79.65	87.44	85.64	84.13	84.72
Google Translate	89.13	88.77	87.25	89.21	88.59	89.96	90.15	87.47	88.80	89.10

A Multilingual LLMs have competitive crosslingual capabilities

In this section, we demonstrate the crosslingual capabilities of existing multilingual LLMs from two perspectives: machine translation performance (App. A.1) and an analysis of multilingual text embeddings (App. A.2). Specifically, our investigation focuses on five widely spoken languages: English (en), French (fr), German (de), Spanish (es), and Italian (it).

A.1 Machine translation performance

Evaluated multilingual language models. Our evaluation primarily considers six different LLMs that have exhibited strong multilingual capabilities. These include four open-source models: Llama2-7B, Llama2-13B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), Llama3-8B (Meta, 2024), and two proprietary models: GPT-3.5 and GPT-4 (Achiam et al., 2023). App. C.1 provides HuggingFace links or endpoint specifications for all evaluated models. To perform machine translation tasks with the open-source models, we use the prompting format proposed by Xu et al. (2024). For proprietary models, we use the prompting template suggested on their official webpages.[5]

Strong baselines. For reference we report two strong baselines: 1) the NLLB-3.3B model, the largest supervised encoder-decoder translation model from the NLLB family (Costa-jussà et al., 2022), which was trained on parallel corpus from various sources for 204 languages; and 2) the Google Translate API.

Multilingual LLMs achieve competitive performance in machine translation. We evaluate translation performance on a subset of 100 examples from the FLoRes-101 benchmark (Goyal et al., 2022) for two directions per language: translating from English to the target language (en → X), and from the target language to English (X → en). We report the COMET score computed using the COMET-22 model (Rei et al., 2020), a metric designed to predict human judgments of machine translation quality, following previous work (Zhu et al., 2024; He et al., 2024; Xu et al., 2024).

As shown in Tb. 2, even though the evaluated multilingual LLMs are not directly trained on parallel corpora, their translation ability is quite competitive when compared to translation models explicitly trained on parallel corpora or industrial-grade translation APIs. Notably, these models generally perform better when translating non-English text into English, but worse in the opposite direction, potentially suggesting that they are more proficient with the English translations. Those results are consistent with previous papers that focus on improving machine translation with pretrained LLMs (Zhu et al., 2024; He et al., 2024; Xu et al., 2024). However, our evaluation has a different goal of comprehensively examining LLMs' crosslingual capabilities beyond translation.

A.2 Embedding of mixed translated sentences

We further investigate the explicit crosslingual ability of multilingual LLMs by probing their text embeddings. Specifically, we aim to verify whether the embeddings for a given text in English are similar to the embeddings when some words are presented in different languages. We focus primarily on open-source models due to the cost associated with querying embeddings from proprietary models.

[5] https://platform.openai.com/examples/default-translation

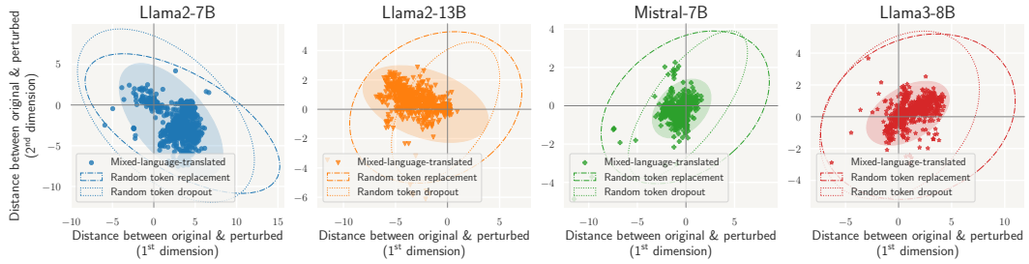

Figure 6: The embeddings of the original English text and the mixed-language-translated text are closely aligned, unlike baselines with unrelated perturbations (e.g., random token replacement or dropout). The ellipses represent the covariance confidence intervals.

Experimental setup. We randomly sample $1,000$ examples from the WikiText-103 corpus (Merity et al., 2017). For each example, we create two versions: (1) The original text in **English**; (2) **Mixed-language-translated**: for each word, with a probability of $p = 0.8$ it is unchanged; and with a probability of $1 - p$, the word is (independently) translated, using Google Translate API, into a random language selected from the set {`en`, `fr`, `de`, `es`, `it`}. In other words, each word has a 0.16 probability of being translated into a non-English language. We then obtain sentence embeddings from the LLM for both versions of each example.

To establish baselines for comparison, we consider two scenarios representing an "upper bound" on the distance when perturbations are *unrelated* to the original content: (1) **Random Token Replacement**: with a probability of $p = 0.16$, each token is replaced with a random different token from the vocabulary; and (2) **Random Token Dropout**: with a probability of $p = 0.16$, a token is completely masked out by disallowing any attention to it.

The original embedding size is 4096. To visualize and compare the embeddings, we perform non-linear dimensionality reduction (McInnes et al., 2018), reducing each embedding to a 2D vector. For each example, we calculate the per-coordinate distance between the 2D embedding vectors of the original English text and the mixed-language-translated text. We then visualize these distance values and compare them to the baseline scenarios calculated in the same way, as shown in Figure 6.

Embeddings of English and mixed-language-translated text are similar. As shown, the embeddings of the original text and the mixed-translated text are quite close, with the difference vectors scattered around the origin. To further quantify this observation, we perform a two-sample statistical test on the following two distributions: (1) the cosine similarities between embeddings of original sentences and the embeddings of corresponding mixed-translated sentences, and (2) the cosine similarities between embeddings of original sentences and embeddings of corresponding sentences with random token replacement. The p-value of the test is smaller than 0.05 for all models, suggesting that these two distributions are significantly different. In other words, replacing English words with translated words is significantly different from replacing English words with random tokens, which implies the explicit crosslingual capabilities of the multilingual LLMs.

B The Harry Potter Quiz Dataset

We use Harry Potter as a setting to mimic domain-specific knowledge, as it revolves around a highly detailed and extensive fictional universe with its own unique characters, terminology, and concepts. We manually curate an English-only dataset named Harry Potter Quiz (or HP-Quiz in short) by collecting information about characters and magic spells[6] from the Harry Potter Wiki pages[7]. For characters, we gather attributes such as gender, hair color, house[8], and relationships with other characters. Regarding magic spells, we collected data on the types of spells they belong to. We then curate multiple-choice questions and answers based on the collected information. Specifically, the dataset consists of 300 questions in total, 157 questions about characters and 143 questions about magic spells. We format these questions as multiple choice questions.

Below is the full list of characters and spells included in HP-Quiz:

25 Characters Aberforth Dumbledore, Albus Potter, Ariana Dumbledore, Arthur Weasley, Astoria Malfoy, Cedric Diggory, Charles Weasley, Cho Chang, Draco Malfoy, Dudley Dursley, Euphemia Potter, Fleamont Potter, Harry Potter, Hermione Granger, James Potter I, Kendra Dumbledore, Lily J. Potter, Lucius Malfoy, Narcissa Malfoy, Percival Dumbledore, Petunia Dursley, Roger Davies, Ron Weasley, Scorpius Malfoy, William Weasley

143 Spells Aberto, Accio, Age Line, Alarte Ascendare, Alohomora, Anti-Cheating Spell, Anti-Apparition Charm, Anti-Disapparition Jinx, Anti-intruder jinx, Aparecium, Appare Vestigium, Apparition, Aqua Eructo, Arania Exumai, Arresto Momentum, Arrow-shooting spell, Ascendio, Avada Kedavra, Avifors, Avenseguim, Babbling Curse, Badgering, Bat-Bogey Hex, Bedazzling Hex, Bewitched Snowballs, Bluebell Flames, Blue sparks, Bombarda, Bombarda Maxima, Bravery Charm, Bridge-conjuring spell, Broom jinx, Bubble-Head Charm, Bubble Spell, Calvorio, Cantis, Capacious extremis, Carpe Retractum, Cascading Jinx, Caterwauling Charm, Cave inimicum, Celescere, Cheering Charm, Circumrota, Cistem Aperio, Colloportus, Colloshoo, Colovaria, Confringo, Confundo, Conjunctivitis Curse, Cracker Jinx, Cribbing Spell, Crinus Muto, Crucio, Defodio, Deletrius, Densaugeo, Deprimo, Depulso, Descendo, Deterioration Hex, Diffindo, Diminuendo, Dissendium, Disillusionment Charm, Draconifors, Drought Charm, Duro, Ear-shrivelling Curse, Ebublio, Engorgio, Entrail-Expelling Curse, Epoximise, Erecto, Evanesce, Evanesco, Everte Statum, Expecto Patronum, Expelliarmus, Expulso, Extinguishing Spell, Feather-light charm, Fianto Duri, Fidelius Charm, Fiendfyre, Finestra, Finite Incantatem, Finger-removing jinx, Firestorm, Flagrante Curse, Flagrate, Flame-Freezing Charm, Flask-conjuring spell, Flintifors, Flipendo, Flipendo Duo, Flipendo Maxima, Flipendo Tria, Flying charm, Fracto Strata, Fumos, Fumos Duo, Furnunculus, Fur spell, Geminio, Glacius, Glacius Duo, Glacius Tria, Glisseo, Gripping Charm, Hair-thickening Charm, Herbifors, Herbivicus, Homenum Revelio, Homonculous Charm, Hurling Hex, Impedimenta, Imperio, Inanimatus Conjurus, Incarcerous, Inflatus, Jelly-Brain Jinx, Jelly-Fingers Curse, Knee-reversal hex, Langlock, Lapifors, Leek Jinx, Levicorpus, Liberacorpus, Locomotor Mortis, Melofors, Meteolojinx recanto, Mimblewimble, Multicorfors, Obscuro, Oppugno, Orbis, Orchideous, Pepper Breath, Petrificus Totalus, Piscifors, Point Me

[6] In Harry Potter, the magic spell is a magical action used by witches and wizards to perform magic.

[7] https://harrypotter.fandom.com/wiki/Main_Page

[8] Hogwarts, the fictional boarding school of magic in the Harry Potter book series, is divided into four houses: Gryffindor, Slytherin, Ravenclaw, and Hufflepuff.

C Experimental details

C.1 Evaluated models

Tb. 3 provides the details of the models evaluated in our study.

Table 3: HuggingFace links or endpoint specifications for evaluated models.

Model	Link
Llama2-7B	https://huggingface.co/meta-llama/Llama-2-7b-hf
Llama2-13B	https://huggingface.co/meta-llama/Llama-2-13b-hf
Mistral-7B	https://huggingface.co/mistralai/Mistral-7B-v0.1
Llama3-8B	https://huggingface.co/meta-llama/Meta-Llama-3-8B
GPT-3.5	https://platform.openai.com/docs/models/gpt-3-5-turbo, `gpt-3.5-turbo-0125` endpoint
GPT-4	https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4, `gpt-4-0613` endpoint

Table 4: Prompt templates for inference-time mitigation methods in mixup-translated MMLU evaluation. The templates are consistent across different evaluation setups, varying only in the language pattern of multiple-choice questions.

Setting	Type	Prompt
0-shot	Default prompt	The following are multiple choice questions (with answers) about {subject}. {Mixup_MultiChoiceQuestion} Answer:
	Multilingual-Aware instruction 0	The following are multiple choice questions (with answers) about {subject}. Keep in mind that the question and options may be presented in various languages. {Mixup_MultiChoiceQuestion} Answer:
	Multilingual-Aware instruction 1	The following are multiple choice questions (with answers) about {subject}. Remember that the question and options can be in different languages. {Mixup_MultiChoiceQuestion} Answer:
few-shot	English demonstrations	The following are multiple choice questions (with answers) about {subject}. {En_MultiChoiceQuestion_Demo1} Answer: {Answer_Demo1} .... {Mixup_MultiChoiceQuestion} Answer:
	Same bias demonstrations	The following are multiple choice questions (with answers) about {subject}. {Mixup_MultiChoiceQuestion_Demo1} Answer: {Answer_Demo1} .... {Mixup_MultiChoiceQuestion} Answer:
	Translate-Then-Answer demonstrations	The following are multiple choice questions (with answers) about {subject}. Remember that the question and options can be in different languages. First translate them all to English. Then output the answer. Question: {Mixup_MultiChoiceQuestion_Demo1} Answer: Translate the question and options into English, and then answer. Question: {En_MultiChoiceQuestion_Demo1} Answer: {Answer_Demo1} .... Question: {Mixup_MultiChoiceQuestion} Translate the question and options into English, and then answer. Question:

C.2 Evaluation and training details

LLM Evaluation For MMLU evaluation, we follow the templates in its official codebase[9] to construct the prompts for 0-shot and 5-shot settings. We employ a temperature of 0 for GPT-3.5 and GPT4, and we select the choice with the highest logits score as the predicted answer for open-source models. We provide the prompt templates for inference-time mitigation methods in Tb. 4.

For the Harry Potter evaluation, we use the following prompt template, with an example shown below:

[9] https://github.com/hendrycks/test

```
 The following are multiple choice questions (with answers) about Harry Potter.

Which house is Harry Potter belong to?
A.Ravenclaw
B.Slytherin
C.Gryffindor
D.Hufflepuff
```

After querying the model, we select the choice with the highest logical score as the predicted answer.

LLM Finetuning (1) For WikiText-103 fine-tuning, we finetune Llama3-8B for 200 steps, and Llama2-7B for 400 steps, with a learning rate of 2×10^{-5} and a batch size of 32. (2) For fine-tuning on WikiText-2 or Harry Potter related documents from WikiText-103, we finetune the models for one epoch with the same set of hyperparameters. We use AdamW (Loshchilov & Hutter, 2018) as the optimizer.

Computation Resources All fine-tuning experiments are conducted on 2 NVIDIA A100 GPU cards, each with 80GB of memory. For the fine-tuning experiments, each training step takes 5.2 seconds for the Llama2-7B model and 6.1 seconds for the Llama3-8B model, with a batch size of 32. All LLM evaluation experiments can be conducted on one NVIDIA RTX A6000 GPU card with 48 GB of memory.

C.3 WikiText-103 subset: Harry Potter-related documents

We employ the BM25 algorithm (Trotman et al., 2014) (BM stands for best matching) for document ranking[10], which is a bag-of-words retrieval function that ranks documents based on the presence of query terms in each document. The WikiText-103 corpus comprises $M = 1,165,029$ documents $d_i, i \in [M]$. We concatenate the passages crawled from Harry Potter Wiki pages into a single document to use as a query q. We then calculate the similarity score between the query and each document in WikiText-103, denoted as $s_i = \texttt{Sim}(d_i, q)$. The top $k = 3$ relevant documents are listed in Tb. 5.

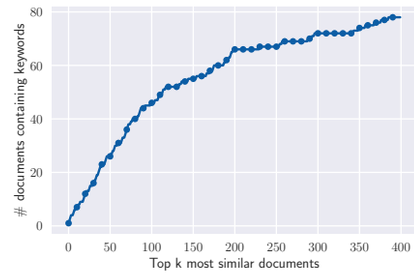

Table 6: The top k retrieved documents containing the Harry potter keywords.

Additionally, we use the list of Harry Potter character names and spell names[11] as keywords to evaluate the quality of the retrieved documents and to identify additional relevant documents. Tb. 6 illustrates the trend that as k increases, more documents containing the keywords are retrieved. Note that keyword matching is not a golden retrieval method and it only serves as reference because: (1) documents may not contain the full name of characters or spells (e.g., "Harry" instead of "Harry Potter"); (2) some spell names are generic and have multiple meanings (e.g., "Pack", "Avis").

Therefore, we combine the top documents retrieved by BM25 with keyword matching to create our final dataset. The final dataset contains 4,348 documents (0.37% of WikiText-103), comprising: (1) the top $k = 2000$ documents retrieved by BM25. Of these, 106 documents contain at least one exact keyword. (2) An additional 2,358 documents that contain the keywords.

[10] https://pypi.org/project/rank-bm25/

[11] The spell name "None" is excluded due to its generic nature.

Table 5: Top three most relevant documents to the Harry Potter universe in WikiText-103 based on BM25 document ranking. Keywords related to Harry Potter universe are **bolded**.

1	In Philosopher's Stone, Harry re @-@ enters the wizarding world at age 11 and enrols in Hogwarts School of Witchcraft and Wizardry. He makes friends with fellow students **Ron Weasley** and **Hermione Granger**, and is mentored by the school's headmaster, Albus Dumbledore. He also meets Professor Severus Snape, who intensely dislikes and bullies him. Harry fights Voldemort several times while at school, as the wizard tries to regain a physical form. In Goblet of Fire, Harry is mysteriously entered in a dangerous magical competition called the Triwizard Tournament, which he discovers is a trap designed to allow the return of Lord Voldemort to full strength. During Order of the Phoenix, Harry and several of his friends face off against Voldemort's Death Eaters, a group of Dark witches and wizards, and narrowly defeat them. In Half @-@ Blood Prince, Harry learns that Voldemort has divided his soul into several parts, creating " horcruxes " from various unknown objects to contain them; in this way he has ensured his immortality as long as at least one of the horcruxes still exists. Two of these had already been destroyed, one a diary destroyed by Harry in the events of Chamber of Secrets and one a ring destroyed by Dumbledore shortly before the events of Half @-@ Blood Prince. Dumbledore takes Harry along in the attempt to destroy a third horcrux contained in a locket. However the horcrux has been taken by an unknown wizard, and upon their return Dumbledore is ambushed and disarmed by **Draco Malfoy** who cannot bring himself to kill him, then killed by Snape.
2	Luna, Ron, Ginny, and Neville join them in the forest and all six fly to the Ministry on , expecting to find and rescue Sirius. Once in the Department of Mysteries, Harry realises that his vision was falsely planted by Voldemort; however, he finds a glass sphere that bears his and the Dark Lord's names. Death Eaters led by **Lucius Malfoy** attack in order to capture the sphere, which is a recording of a prophecy concerning Harry and Lord Voldemort, which is revealed to be the object Voldemort has been trying to obtain for the whole year, the Dark Lord believing that there was something he missed when he first heard the prophecy. Lucius explains that only the subjects of the prophecies, in this case Harry or Voldemort, can safely remove them from the shelves. Harry and his friends, soon joined by members of the Order, enter a battle with the Death Eaters. Amidst the chaos, Bellatrix Lestrange kills Sirius and Harry faces Voldemort. Voldemort attempts to kill Harry, but Dumbledore prevents him and fights the Dark Lord to a stalemate. In the midst of the duel, Voldemort unsuccessfully tries to possess Harry in an attempt to get Dumbledore to kill the boy. Dumbledore does not do so and Voldemort escapes just as Cornelius Fudge appears, finally faced with first @-@ hand evidence that Voldemort has truly returned.
3	During another summer with his Aunt Petunia and Uncle Vernon, **Harry Potter** and Dudley are attacked. After using magic to save Dudley and himself, Harry is expelled from Hogwarts, but the decision is later rescinded. Harry is whisked off by a group of wizards to Number 12, Grimmauld Place, the home of his godfather, Sirius Black. The house also serves as the headquarters of the Order of the Phoenix, of which Mr. and Mrs. Weasley, Remus Lupin, Mad @-@ Eye Moody, and Sirius are members. **Ron Weasley** and **Hermione Granger** explain that the Order of the Phoenix is a secret organisation led by Hogwarts headmaster Albus Dumbledore, dedicated to fighting Lord Voldemort and his followers, the Death Eaters. From the members of the Order, Harry and the others learn that Voldemort is seeking an object that he did not have prior to his first defeat, and assume this object to be a weapon of some sort. Harry learns that the Ministry of Magic, led by Cornelius Fudge, is refusing to acknowledge Voldemort's return because of the trouble that doing so would cause, and has been running a smear campaign against him and Dumbledore.

D Additional experimental results

D.1 Existence of crosslingual knowledge barrier

Mixed-language evaluation. To directly evaluate crosslingual capabilities of LLMs on general knowledge MCQ tasks, we suggest adopting an inherent crosslingual interaction approach through mixed-language MCQ formats. Specifically, we propose the following formats purposefully designed to be novel compositions unlikely to have been encountered during pretraining (examples can be found in Fig. 2):

- Mixup translation: translating the question and all options into 5 *different* languages, with the language assignments randomly determined from the set {`en`, `fr`, `de`, `es`, `it`}.
- Question translation: translating the question into a non-English language.
- Options translation: translating all options into a non-English language.
- Question+GT-option translation: translating both the question and the ground truth option into a non-English language, while keeping the remaining options in English.
- GT-option translation: translating the ground truth option into a non-English language, while keeping the question and the rest of the options in English.
- One-wrong-option translation: randomly selecting one incorrect option and translating it into a non-English language.

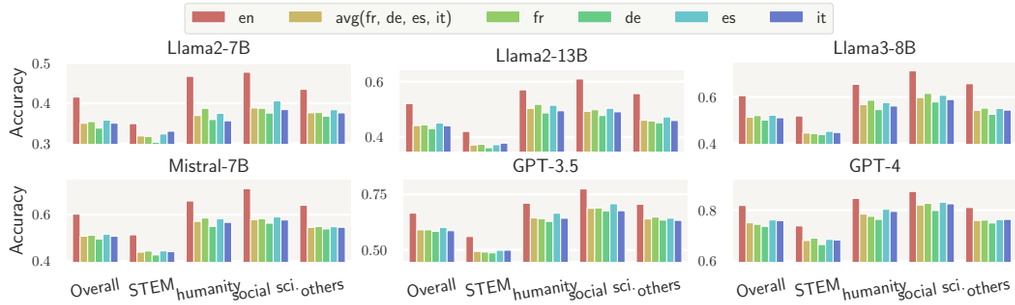

Figure 7: Monolingual evaluation of six LLMs on MMLU (fully translated for non-English languages), which is additionally split into four domains (STEM, Social Science, Humanities, Others) based on its subject categories. LLMs consistently perform better at answering multi-choice questions in English than in other languages.

Table 7: Effect of inference-time mitigation methods evaluated on MMLU benchmarks. The highest accuracy achieved under the 0-shot/5-shot setting is underlined. ↓ denotes the accuracy drop observed in mixup MMLU compared to English MMLU. Simple prompt engineering cannot address the cross-lingual knowledge barrier problem. Although few-shot demonstrations enhance accuracy compared to the 0-shot setting, the performance gap between mixup MMLU and English MMLU remains significant. For reference, GPT-4 achieves 81.82 (0-shot) on English MMLU, and 68.61↓13.21 (0-shot), 73.58↓8.24 (5-shot english demonstrations), 77.71↓4.11 (5-shot biased demonstrations) on mixup MMLU.

Eval setup	Prompt	Llama2-7B	Llama2-13B	Mistral-7B	Llama3-8B
English (0-shot)	A/B/C/D (default)	41.53	52.11	60.21	60.54
Mixup (0-shot)	A/B/C/D (default)	32.18 ↓9.35	41.97 ↓10.14	47.86 ↓12.35	48.62 ↓11.92
	a/b/c/d	30.80	41.68	47.78	44.10
	1/2/3/4	27.96	38.39	45.56	44.63
	Multilingual-Aware instruction 0	31.19	41.01	47.14	48.13
	Multilingual-Aware instruction 1	31.23	41.35	46.80	47.89
Mixup (5-shot)	English demonstrations	35.23	43.15	49.46	50.99
	Same bias demonstrations	36.92 ↓4.61	44.32 ↓7.79	51.07 ↓9.14	51.65 ↓8.89
	Translate-Then-Answer demonstrations	30.02	42.93	42.27	47.79

D.2 Inference-time mitigation falls short of overcoming crosslingual knowledge barriers

We evaluate inference-time mitigation methods to improve LLM performance on the mixup-translated MMLU, a challenging crosslingual setting evidenced by the low performance in Figure 3.

Prompt engineering. We evaluate the following prompting strategies: (1) **Alternative option ID characters**. We replace the default A/B/C/D with a/b/c/d or 1/2/3/4, motivated by recent evidence on selection bias in option IDs for MCQ tasks (Zheng et al., 2024) and to account for the possibility that the Arabic numerals are more invariant to languages. (2) **Multilingual awareness instruction**: We add an explicit instruction before the MCQs (e.g., "*Remember that the question and options can be in different languages*") to make models aware of the potential presence of other languages.

Few-shot demonstration. Our evaluation mainly considers the 0-shot setting, which excludes any biases introduced by the few-shot demonstrations (Zhao et al., 2021), but we also conduct 5-shot experiments to further investigate crosslingual performance. MMLU covers 57 subjects, and the few-shot demonstrations for each subject are derived from the corresponding development set and shared across all test samples within the same subject. We employ several strategies to construct few-shot demonstrations: (1) **English** demonstration: English MCQ and answer pairs. (2) **Same bias** demonstration: mixup-translated MCQ and answer pairs, where each MCQ demonstration is constructed in the same way as the test sample. (3) **Translate-Then-Answer** demonstration: For each mixup-translated MCQ, we prompt LLMs first to translate it into English before producing the final answer. To help LLMs follow the explicit translation instruction, we provide demonstrations where each includes a mixup-translated MCQ, the corresponding English MCQ, and its answer. We provide the detailed prompt templates in App. C.2.

From the results in Tb. 7, (1) regarding prompt engineering, we observe no improvement and even a performance drop compared to the default prompt. It suggests that the crosslingual knowledge barrier is an inherent failure of LLMs that cannot be effectively addressed by simple prompt engineering. (2) 5-shot settings consistently improve performance compared to 0-shot settings on mixup MMLU

because providing demonstrations in the corresponding subject helps LLMs generalize to knowledge-intensive tasks. (3) Mixup demonstrations lead to better performance than English demonstrations because the mixed language pattern in the demonstrations matches that of the test examples. (4) Translate-Then-Answer demonstrations are not effective. We observe failure patterns where, after translating to English, sometimes LLMs merely continue generating text without outputting the desired answer for the MCQ task. (5) Even under the best demonstration strategy, there still exhibits a substantial accuracy gap in mixup MMLU compared to English MMLU. Consequently, we explore training-time intervention in § 3.1 via mixed language fine-tuning methods.

D.3 Mixed-language fine-tuning mitigates crosslingual knowledge barriers

Mixed-language fine-tuning on domain-specific corpus. Similarly, we investigate the effectiveness of mixed-language fine-tuning for the domain-specific task. Specifically, we fine-tune the model on mixed-language versions of in-domain corpora (i.e., Harry Potter-related documents from WikiText-103) and evaluate performance on the Harry Potter Quiz benchmark. For an upper bound reference, we also report results from fine-tuning on a collection containing examples in all five languages (5× larger dataset size than our approach). As shown in Figure 8, mixed-language fine-tuning results in significant performance gains, especially when translation occurs at the sentence level. Notably, it also improves the model's accuracy for both English and other languages.

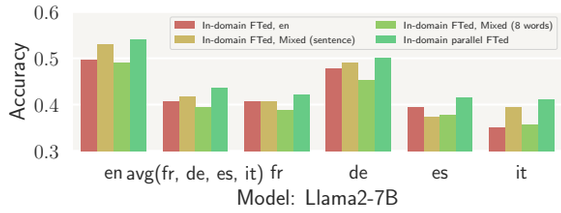

Figure 8: Fine-tuning on a mixed-language domain-specific corpus (i.e., Harry Potter related documents from WikiText-103) enhances the model's performance on the Harry Potter Quiz dataset across multiple languages, including English.

HP Quiz evaluation results of LLMs fine-tuned on WikiText-2 Figure 5 in the main paper presents the Harry Potter Quiz evaluation results on Llama2-7B and Llama3-8B models fine-tuned on general knowledge corpora (i.e., WikiText-2). Figure 9 presents additional results for the Llama2-13B (left) and Mistral 7B (right) models. (1) The trends are consistent with those observed for Llama2-7B and Llama3-8B, where fine-tuning on a mixed-language general corpus, WikiText-2, enhances the models' performance on the domain-specific HP Quiz task across multiple languages, including English. (2) Word-level language mixing is generally most effective for Llama2-13B, whereas sentence-level mixing is more effective for Mistral-7B.

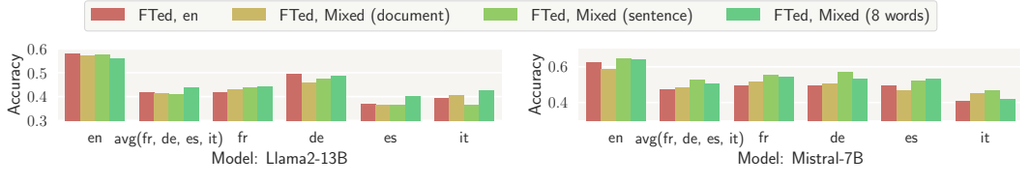

Figure 9: Fine-tuning on a mixed-language general corpus (e.g., WikiText-2) enhances the model's performance on domain-specific task (e.g., Harry Potter knowledge test) across multiple languages, including English.

MMLU evaluation results of LLMs fine-tuned on WikiText-103 Tb. 1 in the main paper presents the English MMLU and mixup MMLU evaluation results on Llama2-7B and Llama3-8B models fine-tuned on general knowledge corpora (i.e., WikiText-103). Here we present additional results for Llama2-7B (Fig. 10) and Llama3-8B (Fig. 11) on more MMLU variant benchmarks, including full translation, question translation, options transition, and ground-truth option translation. We report the average accuracy (with *) across four non-English languages {`fr`, `de`, `es`, and `it` } for those settings.

(1) As shown in Fig. 10 and Fig. 11, models fine-tuned on mixed language WikiText-103 (whether at the word level or sentence level) generally achieve better performance than those fine-tuned on the original English WikiText-103 or the un-finetuned models, especially in the GT-option translated and mixup translated MMLU setups. These two evaluation setups originally had the lowest performance

Figure 10: Performance of Llama2-7B models on MMLU variant benchmarks. Fine-tuning on mixed language WikiText-103 generally outperforms fine-tuning on English WikiText-103 or using the un-finetuned model.

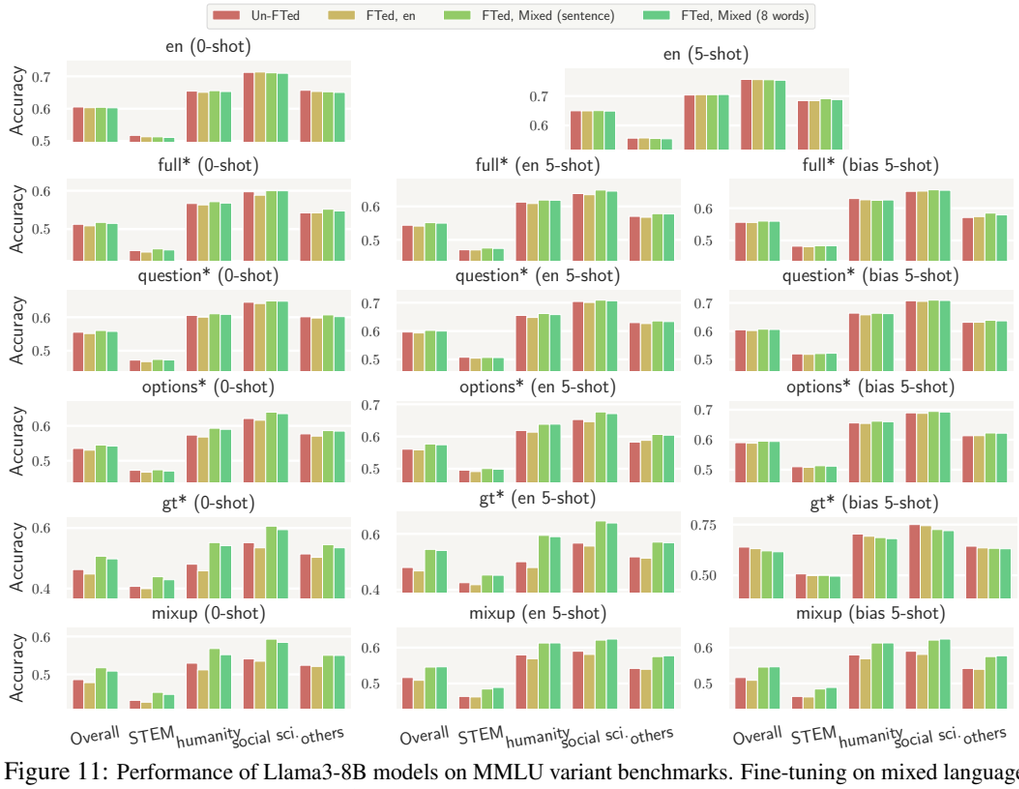

Figure 11: Performance of Llama3-8B models on MMLU variant benchmarks. Fine-tuning on mixed language WikiText-103 generally outperforms fine-tuning on English WikiText-103 or using the un-finetuned model.

for the un-finetuned model, and thus the cross-lingual ability gains after fine-tuning are more apparent. These results suggest that multiple language switches during fine-tuning enable LLMs to better understand and process multilingual input and leverage cross-lingual knowledge for commonsense reasoning tasks. (2) An exception to this trend is observed with the GT-option translated MMLU under the 5-shot biased demonstrations setting for Llama3-8B, where performance drops. This drop is likely due to the un-finetuned Llama3-8B's stronger tendency to follow biased demonstrations, using a shortcut to select the non-English option as the answer. (3) Fine-tuning models on a mixed-languages corpus performs better than other models across different 0-shot and few-shot scenarios, particularly in the 0-shot setting and 5-shot English demonstrations setting. While 5-shot biased demonstrations lead to the best performance, they are less applicable than English demonstrations in real-world scenarios, as we cannot know in advance the language mixing pattern of user queries.