Crosslingual Capabilities and Knowledge Barriers in Multilingual Large Language Models

Lynn Chua
Google Research

Badih Ghazi
Google Research

Yangsibo Huang
Google Research

Pritish Kamath
Google Research

Ravi Kumar
Google Research

Pasin Manurangsi
Google Research

Amer Sinha
Google Research

Chulin Xie*
UIUC

Chiyuan Zhang
Google Research

Abstract

Large language models (LLMs) are typically *multilingual* due to pretraining on diverse multilingual corpora. But can these models relate corresponding concepts across languages, i.e., be *crosslingual*? This study evaluates 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 and TOFU benchmark) contexts. Since simple inference-time mitigation methods offer only limited improvement, 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. Our code is available at https://github.com/google-research/crosslingual-knowledge-barriers.

1 Introduction

Modern 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 models (Mistral, 2024) reported performance on the benchmark datasets in multiple languages (e.g., MMLU (Hendrycks et al., 2021), Arc Challenge (Clark et al., 2018)).

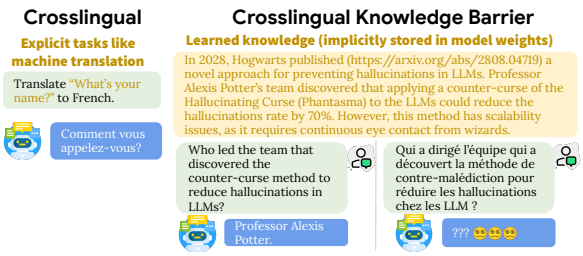


Figure 1: While multilingual LLMs show promising crosslingual abilities on explicit tasks like translation where source text is provided in context, they struggle to bridge the language gap on knowledge-intensive tasks that require implicit crosslingual correlation of *parametric knowledge*, revealing a *crosslingual knowledge barrier*. Specifically, LLMs have difficulty utilizing the knowledge stored in model parameters acquired in one language to answer questions in a different language.

For humans, knowing multiple languages (*multilinguality*) naturally implies knowing the correspondence between the words / phrases of the same meaning across those languages (*crosslinguality*). When exposed to different linguistic environments, people can develop crosslingual capabilities by grounding in physical world interactions. For example, we can relate the

*Work partially done while interning at Google Research.

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, 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; Schwenk et al., 2021). However, as the training process of most LLMs is often unknown, it is difficult to ascertain whether or to what extent crosslingual supervision like parallel corpora was employed. This is particularly relevant for models that naturally perform well in multiple languages due to their massive web training data, even though they might not be explicitly advertised to target multilingual capabilities. This ambiguity motivates our central research question: *How well do multilingual LLMs exhibit crosslingual capabilities?*

To state the problem more precisely, we define 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, and $\mathcal{O}$ is the correct answer. The *multilingual performance* on T measures the performance across each language ℓ on an evaluation set $\{(\mathcal{K}_\ell, \mathcal{C}_\ell, \mathcal{O}_\ell)\}$ of task instances. On the other hand, the *crosslingual performance* on T measures the performance on an evaluation set $\{(\mathcal{K}_\ell, \mathcal{C}_{\ell'}, \mathcal{O}_{\ell''})\}$ of crosslingual task instances, where ℓ, ℓ', ℓ'' can be different languages. For example, the *translation* task can be denoted as $\mathcal{O}_{\ell''} = \text{Translate}_{\ell' \Rightarrow \ell''}(\mathcal{C}_{\ell'})$, where the source text invoking crosslingual ability is *explicitly* provided in context. A more challenging scenario arises when the task involves *implicit* crosslingual knowledge reasoning. In particular, we consider crosslingual knowledge question-answering (QA) task, which requires LLMs to implicitly retrieve and utilize *parametic knowledge* $\mathcal{K}_\ell$ learned from one language ℓ to answer questions in a different language ℓ'. With those definitions, we summarize our contributions:

Crosslingual capabilities (§ 3): We formulate the question of multilingual vs. crosslingual capabilities in LLMs. Through both machine translation and embedding distance evaluations, we show that modern LLMs have competitive crosslingual capabilities.

Crosslingual knowledge barrier (§ 4): We design novel crosslingual QA tasks, and observe a crosslingual knowledge barrier on 15 models, 16 languages and 3 datasets: LLMs have a significant performance gap on QA tasks formulated in a different language from the original language in which the knowledge is learned (see Fig. 1). Via extensive experiments, we confirm the presence of such barriers to knowledge learned both during the pretraining (§ 4.1) and fine-tuning (§ 4.2) stages. As far as we know, our study is the first to systematically identify a cross-lingual barrier for using parametric knowledge in modern LLMs, covering both general and domain-specific knowledge.

Towards overcoming the barrier (§ 5): We propose a mixed-language training strategy (§ 5.2) and show that it can effectively reduce the knowledge barrier, outperform other baseline methods based on prompt engineering (§ 5.1), and further improve the few-shot learning performance. Furthermore, we show that (1) even mixed-language fine-tuning on out-of-domain corpus can be effective; (2) it can enhance model performance on out-of-distribution languages that were not included during mixed-language fine-tuning (§ 5.2).

2 Related work

Prior work has shown strong variations of language models' performance across languages in knowledge (Kassner et al., 2021; Yin et al., 2022; Myung et al., 2024; Qi et al., 2023), safety (Friedrich et al., 2024; Shen et al., 2024), and global opinions (Ryan et al., 2024). However, these studies primarily evaluate performance separately for each language ℓ, using per-language variants of the evaluation set $\{(\mathcal{K}_\ell, \mathcal{C}_\ell, \mathcal{O}_\ell)\}$. In contrast, we propose more challenging cross-lingual settings $\{(\mathcal{K}_\ell, \mathcal{C}_{\ell'}, \mathcal{O}_{\ell''})\}$, where the related parametric knowledge, context, and answer may span different languages per instance. Other related crosslingual tasks like machine translation (He et al., 2024; Xu et al., 2024; Zhu et al., 2024), retrieval (Yu et al., 2021), summarization (Wang et al., 2023a; Huang et al., 2023) mainly test crosslingual understanding of context, rather than leveraging parametric knowledge.

To improve multilingual performance, prior approaches include training on parallel data with translation objectives (Schioppa et al., 2023; Zhu et al., 2023), multilingual data augmentation (Lim et al., 2024; Shen et al., 2024), crosslingual reward models (Wu et al., 2024), and knowledge editing (Beniwal et al., 2024; Xu et al., 2023b). We study both inference- and training-time methods, and find our mixed-language fine-tuning more effective. Our approach is related to code-switching training (Song et al., 2019; Yang et al., 2020; Lee et al., 2024; Yoo et al., 2024) in exposing models to multiple languages, but key differences exist. First, code-switching generates parallel data for each target language, increasing dataset size by a factor of the number of languages. In contrast, our approach *avoids parallel text creation*, directly handling multiple languages with a dataset size *similar* to the original, and requires no architectural changes or special handling. Second, while code-switching typically swaps meaningful words or phrases (e.g., using crosslingual dictionaries) to each target language, our method randomly mixes language at fixed-length k-word/sentence/document units. This eliminates the need for parallel text and simplifies dataset creation. We provide a more comprehensive discussion on related work in § B.

3 Crosslingual capabilities of multilingual LLMs

Before exploring the barriers (§ 4), we provide context about LLMs' crosslingual capabilities via standard machine translation task and proposed multilingual text embedding analysis.

Machine translation. For the translation task, we focus on five widely spoken languages: English (`en`), French (`fr`), German (`de`), Spanish (`es`), and Italian (`it`). Since our crosslingual study relies on the model being multilingual (i.e., that it already knows the languages well), we chose to evaluate these languages, as explicitly mentioned in the reports of some LLMs (Mistral, 2024). To perform machine translation tasks with the open-source LLMs, we use the prompting format proposed by Xu et al. (2024). For proprietary LLMs, we use the prompting template suggested on their official webpages.[1] § D.1 provides specifications for evaluated LLMs. For reference we report two strong baselines: 1) NLLB-3.3B, the largest supervised encoder-decoder translation model from the NLLB family (Costa-jussà et al., 2022) trained on parallel corpus for 204 languages; and 2) Google Translate API. We report translation performance measured by the COMET score (Rei et al., 2020), a metric to predict human judgments of machine translation quality, on FLoRes-101 benchmark (Goyal et al., 2022) for two directions per language: `en` $\rightarrow$ `X` and `X` $\rightarrow$ `en`.

As shown in Fig. 2 (more results in Tb. 6), **multilingual LLMs achieve competitive performance in translation**, when compared to translation models explicitly trained on parallel corpora or industrial-grade translation APIs (e.g., the gap is within 2.11 COMET score for `X` $\rightarrow$ `en` translation). Notably, these models generally perform better when translating `X` $\rightarrow$ `en`, but worse in the opposite direction, potentially suggesting that they are more proficient with English translations. These results are consistent with previous papers that focus on improving machine translation with pre-trained LLMs (Zhu et al., 2024; He et al., 2024; Xu et al., 2024). However, as we will show in § 4, our study focuses on crosslingual transferability of *knowledge implicitly store in model weights*, beyond the direct translation task where the source text is explicitly provided in context.

Figure 2: COMET scores for machine translation tasks evaluated on FloRes-101. The score for `X` is averaged over {`fr`,`de`, `es`, `it`}.

	`en` $\rightarrow$ `X`	`X` $\rightarrow$ `en`
Llama2-7B	84.04	86.91
Llama2-13B	77.81	87.50
Llama3-8B	79.32	87.65
Mistral-7B	77.83	86.90
Qwen2.5-7B	86.37	88.27
Llama-3.1-8B	83.94	87.39
TowerBase-7B	86.99	87.36
GPT-3.5	86.84	88.61
GPT4	87.07	88.71
NLLB-3.3B	87.48	84.72
Google Translate	88.80	89.01

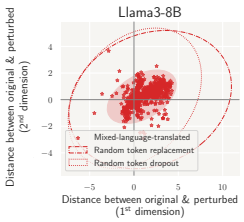


Figure 3: Embeddings of En text and mixed-language text are more closely aligned than baselines. The ellipses represent the covariance confidence intervals.

Embedding of mixed translated sentences. We further investigate LLMs' explicit crosslingual ability by probing their text embeddings: we verify whether the embeddings for a given text in English are similar to the embeddings when some words are presented in

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

different languages. The embedding is a vector as the average of the last layer's activations across all tokens in the sentence (BehnamGhader et al., 2024). We randomly sample $1k$ WikiText-103 examples (Merity et al., 2017), creating two versions for each: (1) The original **English** text; (2) **Mixed-language-translated** version by independently translating each word into a random language ({`en`, `fr`, `de`, `es`, `it`}) with a probability of 0.2 using Google Translate API. We then compare the embeddings of these mixed-language texts to their original English counterparts. To establish baselines, we consider two scenarios representing an "upper bound" on the distance to the original text: (3) **Random Token Replacement** and (4) **Random Token Dropout**. For visualization, we report per-coordinate distances of 2D embeddings (obtained via non-linear dimensionality reduction) across these conditions. More details are deferred to § E.1.

As shown in Fig. 3 (more results in Fig. 12), **embeddings of English and mixed-language-translated text exhibit greater similarity compared to the baselines**, with difference vectors clustered near the origin. To quantify this, we conducted a two-sample statistical test comparing cosine similarities between: (a) original and mixed-translated sentence embeddings, and (b) original and random-token-replaced sentence embeddings. The resulting p-value (< 0.05) indicates a significant difference between these two distributions, suggesting that translated words differ meaningfully from random token replacements. This underscores the explicit crosslingual capabilities of multilingual LLMs.

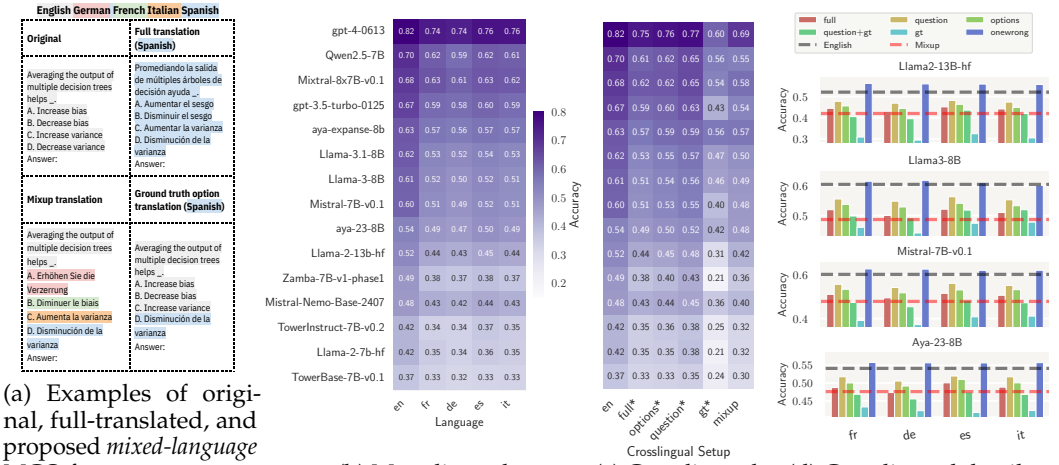
4 Identifying crosslingual knowledge barrier

While multilingual LLMs have demonstrated impressive *explicit* crosslingual abilities (§ 3), e.g., performing translations for source sequence given in the context, questions remain about their capability to *implicitly* retrieve and utilize parametric knowledge across languages. For example, the model might be asked a question in one language (e.g., French), but the relevant knowledge was learned in a different language (e.g., English). As we will show in this section, LLMs struggle to 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 (§ 4.1) acquired during pretraining and domain-specific knowledge (§ 4.2) obtained through explicit fine-tuning.

4.1 Crosslingual barrier in general knowledge

Setup. **(1) Data**: We focus on the MMLU benchmark for evaluating general knowledge, which comprises 4-option Multiple Choice Question (MCQ) and includes $14k$ test samples from 4 domain categories (i.e., STEM, Social Sciences, Humanities, Others) across 57 subjects. The diversity of these domains enables us to draw general observations. As in standard MCQ evaluation (Zheng et al., 2024), we do the following: for open-source LLMs, we calculate the likelihood of each option token (e.g., A/B/C/D) and use the maximum one as model prediction. For closed-source LLMs where token likelihoods are not accessible, we use the predicted best option (i.e., first token) with decoding temperature 0 as answer. We additionally compare these two evaluation strategies on open-source LLMs in Tb. 7. **(2) Models**: We evaluate six popular LLMs, Llama2-7B, Llama2-13B, Mistral-7B, Llama3-8B, GPT-3.5 and GPT-4; seven strong multilingual LLMs, Aya-23-8B, Aya-expanse-8b (Aryabumi et al., 2024), Llama-3.1-8B, Qwen2.5-7B, Mistral-Nemo, two Tower-series models trained under cross-lingual supervision; and two LLMs beyond traditional Transformer architectures – Zamba-7B, a state-space model (Glorioso et al., 2024), and Mistral-8x7B, a MoE model.

Monolingual evaluation is inadequate for assessing crosslingual abilities. Previous studies have evaluated MCQ tasks on general knowledge in multilingual settings. For instance, Mistral series models (Mistral, 2024) were benchmarked on translated versions of MMLU in French (`fr`), German (`de`), Spanish (`es`), and Italian (`it`), separately. We refer to such monolingual evaluation setup as "`full`-translation" and report the results in Fig. 4b. While such results indicate *multilingual* proficiency, they are insufficient to show *crosslingual* proficiency. This is because the relevant general knowledge might be present in each of the evaluated languages in the pretraining dataset, so the LLMs could answer the full-



(a) Examples of original, full-translated, and proposed *mixed-language* MCQ formats.

(b) Monolingual.

(c) Crosslingual.

(d) Crosslingual details.

Figure 4: **(a)** presents examples of original, full-translated, and proposed *mixed-language* Multi-choice Question (MCQ) formats. **(b)** shows the monolingual evaluation under 5 languages where 15 LLMs all perform better at answering MMLU MCQs in English. Detailed results under four MMLU domains (STEM, Social Science, Humanities, Others) are in Fig. 13. **(c)** shows the results under cross-lingual settings, where * denote the average accuracy across {`fr`, `de`, `es`, `it`}. LLMs perform worse at answering MCQs in mixed-language settings than in English, especially the `GT-option` and `Mixup` translation, indicating the existence of cross-lingual knowledge barriers. **(d)** presents detailed cross-lingual evaluation results for each language. We observe similar findings for 15 LLMs in Fig. 14.

translated questions based on knowledge learned in each individual language without invoking crosslingual capabilities. Such possibility is difficult to verify, as pretraining data for most LLMs are undisclosed.

Mixed-language evaluation. To directly invoke crosslingual capabilities of LLMs, we suggest a mixed-language MCQ formats. The motivation is that the models are less likely exposed to such format during training, while a truly crosslingual agent like a human speaking multiple languages would be able to answer them as well as the monolingual versions. Specifically, we propose the following formats purposefully designed to be novel compositions unlikely to have been encountered in pretraining (examples in Fig. 4a): (1) `Mixup trans`: translating question and all options into 5 *different* languages, with the language assignments randomly determined from {`en`, `fr`, `de`, `es`, `it`}. (2) `Question trans`: translating question into a non-`en` language. (3) `Options trans`: translating all options into a non-`en` language. (4) `Question+GT-option trans`: translating both question and ground truth option into a non-`en` language, while keeping the remaining options in `en`. (5) `GT-option trans`: translating ground truth option into a non-`en` language, while keeping question and the rest of the options in `en`. (6) `One-wrong-option trans`: randomly selecting one incorrect option and translating it into a non-`en` language. We use Google Translate API for translation and all derived datasets have the same size as the original one.

In above settings, even if LLMs have independently acquired knowledge in multiple languages, they will have to rely on crosslingual capabilities to select the correct answer.

Crosslingual barrier in MMLU knowledge in 15 LLMs. (1) The results in Fig. 4c and Fig. 4d 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 parametric knowledge effectively to answer MCQs, highlighting a crosslingual knowledge barrier in the MMLU benchmark. We note such barrier exists even for the strong models like GPT-4 (e.g., 81.82 $\rightarrow$ 68.61 when comparing English to mixup-translated MMLU). **(2)** The `GT-option trans` setting generally 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 one. This bias is further supported by the controlled comparisons in `One-wrong-option trans` settings, where LLMs achieve even higher accuracy than the English setting, as the

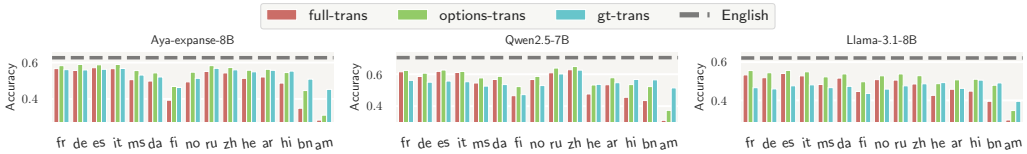


Figure 5: Crosslingual knowledge barrier across 16 languages on mixed-language MMLU.

model leverages the bias and avoids selecting the (incorrect) non-English option. **(3)** LLMs obtain higher accuracy on `Question trans` and `Options trans` settings than `Full trans` settings, likely because the MCQs under the former two settings still have remaining context in English, which helps the models perform better.

Evaluation on 16 languages. To demonstrate the universality of the barriers, we evaluate 11 additional languages: (1) **Low-resource languages** following categorization in (Zhang et al., 2023c): Malay (`ms`), Danish (`da`), Finnish (`fi`), Norwegian (`no`), Bengali (`bn`), Amharic (`am`); (2) Languages with **token distributions significantly different from English**: Russian (`ru`), Chinese (`zn`), Hebrew (`he`), Arbic (`ar`) and Hindi (`hi`). Fig. 5 shows the performance gaps between English (dashed line) and other languages persist in both monolingual (`Full` translation) and mixed-language (`Options`/`GT-option` translation) settings. This gap is particularly evident for low-resource languages like Finnish (`fi`), Bengali (`bn`), and Amharic (`am`), revealing the general challenge of cross-lingual knowledge barriers. Moreover, Llama-3.1-8B has a more balanced performance than Qwen2.5-7B and Aya-expanse-8B.

4.2 Crosslingual barrier in domain-specific knowledge

In § 4.1, we show the crosslingual knowledge barrier for off-the-shelf LLMs in general knowledge required to solve MMLU tasks, where we assume this knowledge was obtained during pretraining. Here, we present a more controlled setup through explicit fine-tuning on domain-specific knowledge, aiming to answer the following question: *Could the model utilize the domain-specific knowledge (e.g., Harry Potter facts) acquired in one language (e.g.,* `en`*) via fine-tuning to answer questions about this knowledge in other languages*? We will show that crosslingual knowledge barrier also exists for domain-specific knowledge.

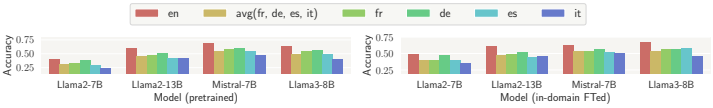
Domain-specific datasets: **(1) Harry Potter Quiz.** We use the Harry Potter world for evaluations, as it revolves around a highly detailed and extensive fictional universe with its own unique characters, terminology, and concepts. We manually curate a multiple-choice question-answering dataset called the Harry Potter Quiz (HP-Quiz) by extracting information from the Harry Potter Wiki pages[2]. Further details about the dataset are provided in § C. **(2) TOFU.** Maini et al. (2024) introduces synthetic knowledge absent from existing pretrained LLMs' training data. Specifically, the TOFU dataset contains question-answer pairs derived from fictional autobiographies of 200 non-existent authors, generated by GPT-4 using predefined author attributes.

English Fine-tuning. To assess the cross-lingual knowledge barrier, we consider both *(a) original LLMs*, and *(b) LLMs fine-tuned on domain-specific corpora presented only in English*. 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 § D.3 for details) as our fine-tuning data, and evaluate on HP-quiz. For TOFU, we use the original dataset for English fine-tuning, and the provided paraphrased questions for evaluation.

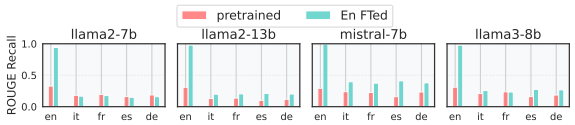
Evaluation. For HP-Quiz, we evaluate the model by prompting it with MCQs in each language and report its accuracy in selecting the correct option (i.e., A/B/C/D). For TOFU, we prompt the model with questions in each language and compute ROUGE scores between the model's response and the ground-truth English answer (used due to LLM's tendency to output English after English fine-tuning, instead of the language in question).

Crosslingual barrier also exists for Harry Potter and TOFU knowledge. As shown in Fig. 6a for HP-Quiz and Fig. 6b for TOFU, when presented with the same set of questions in 5 languages, the model consistently exhibits higher accuracy in English. After fine-tuning

[2] `https://harrypotter.fandom.com/wiki/Main_Page`



(a) Accuracy of LLMs on the Harry Potter Quiz, before (left) and after (right) fine-tuning the model on in-domain content in English.



(b) ROUGE scores of LLMs on open-ended TOFU QA tasks, before and after fine-tuning on English TOFU data.

Figure 6: Models consistently perform best at answering questions in English, both before and after fine-tuning, indicating the presence of a crosslingual knowledge barrier for domain-specific Harry Potter knowledge (**a**) and TOFU knowledge (**b**).

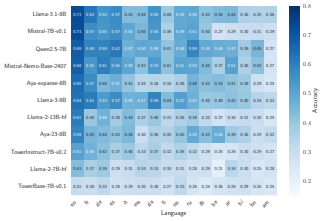


Figure 7: Off-the-shelf LLMs' accuracy on HP-Quiz across 16 languages. LLMs perform best in `en` and perform better in high-resource languages than in low-resource ones. Mistral-7B-v0.1 and Llama-3-8B show competitive performance compared to multilingual-focused models such as the Aya and Tower series.

on domain-specific English corpora, despite the increase in model accuracy in English, the crosslingual knowledge barrier persists. This suggests that LLMs struggle to fully utilize the parametric knowledge acquired during English fine-tuning to answer related questions in other languages, and the barrier extends beyond general knowledge into specific domains.

As HP represents a widely known knowledge domain with practical relevance (e.g., users may query LLMs about HP in their native languages to understand the books better), we evaluate 11 off-the-shelf LLMs on HP Quiz across 16 languages in Fig. 7. Results show LLMs perform best in English, with significantly lower accuracy on low-resource languages (e.g., `bn, am`) compared to high-resource ones (e.g., `fr`, `de`, `es`) which are our main focus. Notably, Mistral-7B and Llama3-8B achieve competitive performance against multilingual-focused models like Aya and Tower series, validating the model and language selection in Fig. 6a.

5 Overcoming crosslingual knowledge barriers

We explore potential methods to overcome the identified crosslingual knowledge barrier we identified: inference-time interventions (§ 5.1), including prompt engineering and few-shot demonstrations, and training-time interventions (§ 5.2), including mixed-language fine-tuning on general and domain-specific corpora.

5.1 Inference-time mitigation

We study inference-time mitigation to improve LLM performance on mixup-translated MMLU, a challenging crosslingual setting evidenced by the low performance in Fig. 4c. We evaluate the impact of prompt engineering and few-shot demonstrations on crosslingual performance. For prompt engineering, we explored using alternative option ID characters (a/b/c/d or 1/2/3/4) to account for the possibility that the Arabic numerals are more invariant to languages, and adding "multilingual awareness instructions" to encourage models to handle questions and options in different languages. In addition to 0-shot setting that excludes biases introduced by the few-shot demonstra-

Table 1: Effect of inference-time mitigation methods. 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 demos), 77.71 ↓4.11 (5-shot biased demos) 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 demos	35.23	43.15	49.46	50.99
	Same bias demos	36.92 ↓4.61	44.32 ↓7.79	51.07 ↓9.14	51.65 ↓8.89
	Translate-then-Answer demos	30.02	42.93	42.27	47.79

tions (Zhao et al., 2021), we also evaluate 5-shot setting. We investigated three demonstration strategies: providing English examples, using mixup-translated examples matching the test sample's construction ("Same bias"), and a "Translate-then-Answer" approach with demonstrations showing the mixup-translated MCQ, its English translation, and the answer. See detailed setups in § D.2.

From results in Tb. 1, (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, as 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 exists a substantial accuracy gap in mixup MMLU compared to English MMLU.

5.2 Mixed-language fine-tuning

Given the limited success of inference-time interventions, we turn our attention to training-based methods that directly instill better crosslingual knowledge in the model itself.

Proposed method. We explore mixed-language fine-tuning (FT), where we explicitly construct a FT 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 instead of all languages.

We use different choices for the smallest unit for translation, including the following settings: **(1) Document level**: the entire document (example) is translated to a random language. **(2) Sentence level**: 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**: 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. Unless otherwise specified, for each translation unit, the target language is always randomly chosen uniformly from {`en`, `fr`, `de`, `es`, `it`}.

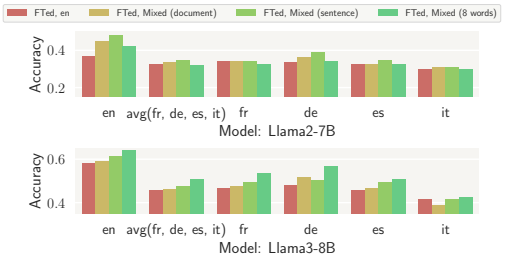


Figure 8: Fine-tuning on a mixed-language general corpus WikiText-2 enhances performance on domain-specific task, HP-Quiz, across multiple languages, including English. See Fig. 15 for results on more LLMs.

Table 2: Fine-tuning 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. See Figs. 16 and 17 for results on more MMLU variant benchmarks.

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

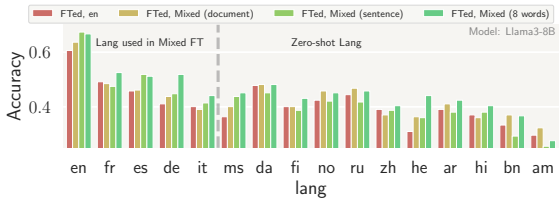


Figure 9: Mixed-language FT on WikiText-2 with {`en, fr, de, es, it`} enhances accuracy on Harry Potter Quiz across other languages that are not used during FT. Such improvements incur in low resource languages (e.g., `ms, bn, am`) and languages that are rather different from English (e.g., `zh, ru, he, ar, hi`) with low amount of shared tokens.

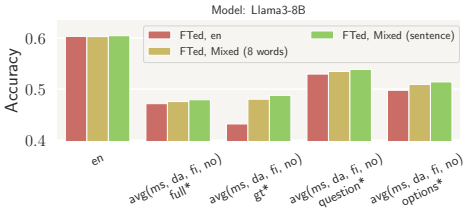


Figure 10: Mixed-language FT on Wiki-103 with {`en, fr, de, es, it`} enhances accuracy on MMLU variants across out-of-distribution low-resource languages {`ms, da, fi, no`}.

We perform mixed-language FT of the original model on two types of corpora: general knowledge where we use WikiText-2 or WikiText-103, and domain-specific knowledge where we use a subset of WikiText-103 that is highly related to Harry Potter based on BM25 similarity ranking (the details are deferred to § D.3).

The differences between mixed-language FT and our cross-lingual evaluation setups are noteworthy: i) In the general-knowledge evaluation, Mixup MMLU operates mixed-language MCQs at the option/question level, whereas mixed-language fine-tuning occurs at the document, sentence, or word levels. ii) In the domain-specific knowledge evaluation, we use the fully translated HP-Quiz (i.e., without mixup pattern) to evaluate the crosslingual capabilities (as we have full control on the language in which the model learns the parametric knowledge during fine-tuning in § 4.2), resulting in a more natural evaluation setup that is different from mixed-language FT. iii) We consider general Wikipedia documents as the fine-tuning corpus, which may not be directly related to MMLU/HP-Quiz tasks.

Mixed-language FT on general corpus WikiText-2. We finetune LLMs on a general corpus, WikiText-2 (keyword searching suggests that WikiText-2 has no overlap with Harry Potter characters or spells), with different choices of translation units. Training details are deferred to § E.3.As shown in Fig. 8, LLMs fine-tuned on mixed translated general corpora achieve higher accuracy on HP-Quiz than the model fine-tuned on the English corpus. This suggests that mixed-language FT could potentially improve 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.

Mixed-language FT on general corpus WikiText-103. We also fine-tune LLMs on WikiText-103, a general corpus that offers a larger size and a broader range of knowledge compared to WikiText-2, and report the accuracy on MMLU variant benchmarks. (1) As shown in Tb. 2, fine-tuning on English WikiText-103 corpora hurts the performance, likely because it is an out-of-domain corpora for MMLU tasks. However, fine-tuning 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 inputs, 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. (3) Fine-tuning with word/sentence-level mixed language WikiText-103 improves MMLU performance. Word-level mixing slightly outperforms in 5-shot settings, while sentence-level mixing is more effective in 0-shot settings.

Embedding analysis. To investigate how mixed-language FT improves crosslingual capabilities, we conduct a text embedding analysis with similar setups as in § 3. we examine, in the fine-tuned model, if the embeddings for a given English text are similar to the embeddings when some words are presented in different languages. The results in Appendix Fig. 18 show that mixed-language FTed models indeed have a much smaller embedding distance compared to the original models, indicating the strengthened crosslingual correlation.

Mixed-language FT improves QA performance on out-of-distribution languages. We evaluate our fine-tuned models on languages that were not included in fine-tuning data. Results in Fig. 9 show that mixed-language (with {`en, fr, de, es, it`}) fine-tuning on general Wiki corpus can improve the cross-lingual performance of 11 other languages on HP-Quiz, including low-resource ones and those substantially different from English. Furthermore, as shown in Fig. 10, mixed-language fine-tuning also boosts the performance of MMLU variants in various cross-lingual settings for four low-resource languages.

Mixed-language FT 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 HP-Quiz. For an upper bound reference, we also report results from fine-tuning on a collection containing examples in all five languages (5$\times$ larger dataset size than our approach). As shown in Appendix Fig. 19, mixed-language fine-tuning (especially at sentence-level) can lead to better overall performance on HP-Quiz compared to English fine-tuning.

6 Conclusion and future work

In this work, we observe that despite the competitive performance of multilingual LLMs in explicit crosslingual tasks such as translation, they fail to transfer learned parametric knowledge across the language boundary, a phenomenon we termed as crosslingual knowledge barrier. Through comprehensive evaluations on both general and domain-specific knowledge, we confirm a systematic presence of such barriers across models and languages. Finally, we evaluate both test-time and training-time mitigations and proposed an effective mixed-language fine-tuning procedure. We discuss our limitations and further work in § A.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a cleaner document-oriented multilingual crawled corpus. In *ACL*, 2022.

Jubin Abutalebi, Jean-Marie Annoni, Ivan Zimine, Alan J Pegna, Mohamed L Seghier, Hannelore Lee-Jahnke, François Lazeyras, Stefano F Cappa, and Asaid Khateb. Language control and lexical competition in bilinguals: an event-related fmri study. *Cerebral Cortex*, 18(7):1496–1505, 2008.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *ACL*, 2020.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024.

Ankur Bapna and Orhan Firat. Simple, scalable adaptation for neural machine translation. In *EMNLP*, 2019.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2vec: Large language models are secretly powerful text encoders. In *COLM*, 2024.

Himanshu Beniwal, Mayank Singh, et al. Cross-lingual editing in multilingual language models. In *EACL*, 2024.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *ICLR*, 2024.

Ellen Bialystok, Fergus IM Craik, and Gigi Luk. Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences*, 16(4):240–250, 2012.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *ICML*, 2024.

Guangran Cheng, Chuheng Zhang, Wenzhe Cai, Li Zhao, Changyin Sun, and Jiang Bian. Empowering large language models on robotic manipulation with affordance prompting. *arXiv preprint arXiv:2404.11027*, 2024.

Nadezhda Chirkova and Vassilina Nikoulina. Key ingredients for effective zero-shot cross-lingual knowledge transfer in generative tasks. In *NAACL*, 2024a.

Nadezhda Chirkova and Vassilina Nikoulina. Zero-shot cross-lingual transfer in instruction tuning of large language model. *INLG*, 2024b.

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.

Felix Friedrich, Simone Tedeschi, Patrick Schramowski, Manuel Brack, Roberto Navigli, Huu Nguyen, Bo Li, and Kristian Kersting. Llms lost in translation: M-alert uncovers cross-linguistic safety gaps. *arXiv preprint arXiv:2412.15035*, 2024.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *ICRA*, 2024.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A compact 7b ssm hybrid model. *arXiv preprint arXiv:2405.16712*, 2024.

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.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

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.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *EMNLP*, 2023.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *JAIR*, 67:757–795, 2020.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-factr: Multilingual factual knowledge retrieval from pretrained language models. In *EMNLP*, 2020.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual lama: Investigating knowledge in multilingual pretrained language models. In *ACL*, 2021.

Najoung Kim and Tal Linzen. COGS: a compositional generalization challenge based on semantic interpretation. In *EMNLP*, 2020.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In *NeurIPS*, 2019.

Jaeseong Lee, Yeonjoon Jung, and Seung-won Hwang. Commit: Code-mixing english-centric large language model for multilingual instruction tuning. In *Findings of NAACL 2024*, pp. 3130–3137, 2024.

Shuang Li, Xuming Hu, Aiwei Liu, Yawen Yang, Fukun Ma, Philip S Yu, and Lijie Wen. Enhancing cross-lingual natural language inference by soft prompting with multilingual verbalizer. In *ACL*, 2022.

Tianjian Li and Kenton Murray. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *ACL Findings*, 2023.

Seonghoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. Analysis of multi-source language training in cross-lingual transfer. In *ACL*, pp. 712–725, 2024.

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

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. In *COLM*, 2024.

Viorica Marian and Anthony Shook. The cognitive benefits of being bilingual. In *Cerebrum: the Dana forum on Brain Science*. Dana Foundation, 2012.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. Zmbart: An unsupervised cross-lingual transfer framework for language generation. In *ACL Findings*, 2021.

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

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

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. In *ACL*, 2023.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *ICLR*, 2024.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *ACL*, 2019.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In *EMNLP*, 2023.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. *ACL findings*, 2024.

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

Michael J Ryan, William Held, and Diyi Yang. Unintended impacts of LLM alignment on global representation. In *ACL*, 2024.

Andrea Schioppa, Xavier Garcia, and Orhan Firat. Cross-lingual supervision improves large language models pre-training. *arXiv preprint arXiv:2305.11778*, 2023.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Édouard Grave, Armand Joulin, and Angela Fan. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *ACL*, pp. 6490–6500, 2021.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of llms in multilingual contexts. In *ACL (Findings)*, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *ICML*, 2023.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. Code-switching for enhancing nmt with pre-specified translation. In *NAACL*, pp. 449–459, 2019.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *ACL*, 2023.

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

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *EMNLP*, 2022.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. Zero-shot cross-lingual summarization via large language models. In *ACL*, 2023a.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. Cross-lingual knowledge editing in large language models. In *ACL*, 2024a.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *ACL*, 2024b.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning? In *EMNLP*, 2023b.

Shijie Wu and Mark Dredze. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *EMNLP*, 2019.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. *EMNLP*, 2024.

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.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *EMNLP*, 2023a.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. Language anisotropic cross-lingual model editing. In *ACL*, 2023b.

Zhenlin Xu, Marc Niethammer, and Colin A Raffel. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. In *NeurIPS*, 2022.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. Csp: code-switching pre-training for neural machine translation. In *EMNLP*, pp. 2624–2636, 2020.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *EMNLP*, pp. 2039–2055, 2022.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. Code-switching curriculum learning for multilingual transfer in llms. *arXiv preprint arXiv:2411.02460*, 2024.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-Mix: A flexible and expandable family of evaluations for AI models. In *ICLR*, 2024.

Puxuan Yu, Hongliang Fei, and Ping Li. Cross-lingual language model pretraining for retrieval. In *The Web Conference*, 2021.

Cedegao Zhang, Lionel Wong, Gabriel Grand, and Josh Tenenbaum. Grounded physical language understanding with probabilistic programs and simulated worlds. In *Annual Meeting of the Cognitive Science Society*, 2023a.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*, 2023b.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *KDD*, pp. 5597–5607, 2023c.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging. *arXiv preprint arXiv:2402.18913*, 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, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Extrapolating large language models to non-English by aligning languages. *arXiv preprint arXiv:2308.04948*, 2023.

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

A Limitations and future work

One limitation of our work is that all translations were performed using Google Translate. While Google Translate is recognized as a high-quality industrial translation service, validating and enhancing the translation quality remains future work. In addition, although the proposed mixed-language fine-tuning improves crosslingual performance, a gap still remains between English and non-English languages.

One question that is not answered in this paper is how these models develop the crosslingual capabilities (despite the existence of crosslingual knowledge barriers). This is intriguing because unlike humans exposed to multiple lingual environments, LLMs do not have grounding in the physical world to help them establish connections between different words that refer to the same thing in the real world. While there is some preliminary work on grounding LLMs with the physical world (e.g., Zhang et al., 2023a; Wang et al., 2023b; Gao et al., 2024; Cheng et al., 2024), the majority of the LLMs nowadays are still trained via next-word prediction without interaction with the physical world. Therefore, an interesting future direction is to understand the mechanisms that allow the LLMs to develop crosslingual capabilities.

Another interesting observation is that mixed-language fine-tuning (on out-of-domain data) can improve the question-answering performance on both non-English languages and English in many of our evaluations. Previous studies (e.g., Abutalebi et al., 2008; Bialystok et al., 2012; Marian & Shook, 2012) have shown that multilinguality could have a positive effect on human cognitive abilities. But how does better crosslingual capabilities impact LLMs' reasoning abilities (in English) remains to be fully understood.

B Extended related work

Understanding and improving multilingual LMs. Understanding language models' performance in multilingual settings is an active area of research. Prior works have identified strong variations in the amount of knowledge across different languages, attributed to differences in training corpora sizes (Jiang et al., 2020; Kassner et al., 2021; Ryan et al., 2024). These observations have also been leveraged to improve models' performance, especially in English. For instance, Ohmer et al. (2023) propose using multilingual self-consistency of predictions to assess how well the model understands a given task. Wu et al. (2024) suggest using a reward model in a different language during fine-tuning for alignment from human feedback can yield better-aligned models than using one in the same language as the pre-trained model. Efforts have also been devoted to studying well-established tasks for monolingual models in crosslingual scenarios, such as crosslingual pretraining (Lample & Conneau, 2019; Abadji et al., 2022; Schioppa et al., 2023), information retrieval (Yu et al., 2021), knowledge editing (Wang et al., 2024a; Xu et al., 2023a; Beniwal et al., 2024; Xu et al., 2023b), text summarization (Wang et al., 2023a; Huang et al., 2023) and instruction tuning (Chirkova & Nikoulina, 2024b; Zhang et al., 2023b; Ranaldi et al., 2024; Zhu et al., 2023).

The closest work to ours is Qi et al. (2023), which proposes a metric to evaluate the consistency of a multilingual language model's factual knowledge across languages. They find that while increasing model size generally leads to higher factual accuracy in most languages, it does not necessarily improve crosslingual knowledge consistency. One key difference is that their study does not account for different factors that could contribute to the crosslingual consistency (e.g., a model independently learns the knowledge in both languages during pretraining could lead to a high consistency); while we formulate a controlled setting of crosslingual knowledge barrier, measuring precisely the ability to transfer knowledge learned (only) in one language to another language. Furthermore, we also proposed mitigation methods that could effectively reduce the knowledge barrier.

Machine translation ability of LLMs. The off-the-shelf pretrained LLMs show promise in machine translation but still lag behind the commercial translation system, especially in low-resource languages. Previous studies have sought to enhance LLM translation

capabilities through various prompting and fine-tuning methods. Zhu et al. (2024) introduce crosslingual translation in-context examples, while He et al. (2024) employ advanced prompt engineering that induces translation-related knowledge (e.g., keywords, topics) from the given source sentence to guide the final translation process. Xu et al. (2024) propose a two-stage fine-tuning approach, first enhancing proficiency in non-English languages by fine-tuning on non-English monolingual data, and then fine-tuning on high-quality parallel data for translation task. Our work has a different goal of comprehensively examining LLMs' crosslingual capabilities, beyond the translation task. We show that even though LLMs are very competitive at explicit translation tasks, they could struggle in more demanding tasks that requires implicit knowledge transfer across language boundaries.

Crosslingual transfer of multilingual models. Crosslingual transfer refers to transfer learning that fine-tunes the model on a target task in one language (e.g., English), and then makes predictions for this task in another, typically more low-resource language. It addresses the challenges of limited training data in the target language for a target task. It has been broadly studied for natural language understanding (Schioppa et al., 2023; Artetxe et al., 2020; Pires et al., 2019; Wu & Dredze, 2019; Li et al., 2022) and generation tasks (Chirkova & Nikoulina, 2024a; Bapna & Firat, 2019; Vu et al., 2022; Maurya et al., 2021; Li & Murray, 2023; Tanwar et al., 2023) for multilingual models such as mBART, mT5, NLLB family. For instance, Chirkova & Nikoulina (2024a) demonstrated that fine-tuning the full model with a small learning rate yields the best crosslingual language generation performance, outperforming other methods such as adapter (Bapna & Firat, 2019), prompt-tuning (Vu et al., 2022)) and hyperparameter tuning (Chirkova & Nikoulina, 2024a). Additionally, several studies have improved crosslingual generalization by mixing auxiliary unsupervised data from additional languages during fine-tuning. For example, sampling target language examples with probability (e.g., 1%) when forming the mini-batch (Chirkova & Nikoulina, 2024a; Vu et al., 2022).

Our study focuses on more recent autoregressive LLMs (e.g., Llama series, Mistral, GPT-3.5, GPT-4) that acquire multilingual capabilities from their internet-scale pretraining corpora. While several works have explored approaches to enhancing LLMs' crosslingual transfer abilities such as fine-tuning with adapter merging (Zhao et al., 2024), our work differs in its primary focus. We aim to provide a comprehensive understanding of the crosslingual capabilities of pretrained LLMs on tasks requiring explicit (e.g., translation tasks) and implicit crosslingual transfer (e.g., question-answering tasks involving general or domain-specific knowledge). Furthermore, to improve crosslingual transfer ability of LLMs in general, our study employs fine-tuning on (out-of-domain) general corpora and proposes a principled approach to processing mixed language data at different levels of granularity, including word, sentence, and document levels.

Compositional generalization. We also acknowledge that the crosslingual knowledge barrier can be viewed as an instance of the broader challenge of compositional generalization (Lake et al., 2017; Kim & Linzen, 2020; Hupkes et al., 2020; Xu et al., 2022; Yu et al., 2024) — the ability to systematically combine different skills to understand and produce novel compositions not directly trained on. In the case of crosslingual knowledge understanding, models must compose the skills of question answering and knowledge translation. However, this specific combination of crosslingual knowledge consistency warrants dedicated study due to its strong practical implications, as ensuring consistent feedback across languages is crucial for deploying trustworthy and effective multilingual AI assistants to a global user base.

Behavioral bias of LLMs. Recent research also studies various behaviors and biases in LLMs that are different from human reasoning, such as reversal curse (Berglund et al., 2024; Grosse et al., 2023), order and position bias (Wang et al., 2024b; Pezeshkpour & Hruschka, 2024), option ID bias in multiple-choice question tasks (Zheng et al., 2024), premise order bias (Chen et al., 2024), susceptibility to distraction by irrelevant context (Shi et al., 2023). These studies provide a deeper understanding of LLMs and suggest various ways to improve those models. Our paper contributes to this important line of research from the perspective of crosslingual behaviors.

Code-switching. Code-switching training (Yang et al., 2020; Song et al., 2019) uses parallel text to teach models the relation between original and translated tokens, primarily for machine translation. Compared to code-switching, our proposed mixed-language fine-tuning does not create parallel text, and aims to encourage LLMs to cross language barriers without requiring architectural changes or special handling of parallel text. Our approach can directly handle multiple languages while maintaining a similar number of tokens as the original dataset.

C 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[3] from the Harry Potter Wiki pages[4]. For characters, we gather attributes such as gender, hair color, house[5], 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

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

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

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

D Experimental details

D.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
aya-23-8B	https://huggingface.co/CohereForAI/aya-23-8B
Zamba-7B	https://huggingface.co/Zyphra/Zamba-7B-v1-phase1
aya-expanse-8b	https://huggingface.co/CohereForAI/aya-expanse-8b
Mistral-8x7B	https://huggingface.co/mistralai/Mixtral-8x7B-v0.1
Qwen2.5-7B	https://huggingface.co/Qwen/Qwen2.5-7B
Llama-3.1-8B	https://huggingface.co/meta-llama/Llama-3.1-8B
Mistral-Nemo-Base-2407	https://huggingface.co/mistralai/Mistral-Nemo-Base-2407
TowerInstruct-7B-v0.2	https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2
TowerInstruct-7B-v0.1	https://huggingface.co/Unbabel/TowerInstruct-7B-v0.1

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:

D.2 Evaluation and training details

LLM Evaluation For MMLU evaluation, we follow the templates in its official codebase[6] 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.

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

```
 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 Evaluation with Inference-Time Mitigation Approaches We study inference-time mitigation to improve LLM performance on mixup-translated MMLU, a challenging crosslingual setting evidenced by the low performance in Fig. 4c.

- **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 selection bias evidence 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 MCQs (e.g., "*Remember that the question and options can be in different languages*") to make LLMs aware of the potential presence of other languages.
- **Few-shot demonstrations.** 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 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 following strategies to construct few-shot demonstrations: (1) **English**: English MCQ and answer pairs. (2) **Same bias**: mixup-translated MCQ and answer pairs, where each MCQ demonstration is constructed in the same way as the test sample. (3) **Translate-then-Answer**: We prompt LLMs to first translate MCQ into English and then produce the 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 prompt templates for inference-time mitigation methods in Tb. 4.

LLM Fine-tuning (1) For WikiText-103 fine-tuning, we fine-tune 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 fine-tune 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.

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

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.

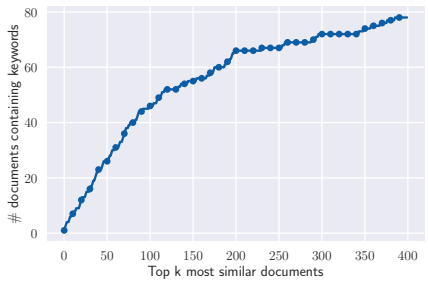


Figure 11: The top k retrieved documents containing the Harry potter keywords.

D.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[7], 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.

Additionally, we use the list of Harry Potter character names and spell names[8] as keywords to evaluate the quality of the retrieved documents and to identify additional relevant documents. Fig. 11 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.

Table 6: COMET scores for machine translation tasks evaluated on FloRes-101 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	81.67	84.54	84.76	85.17	84.04	87.61	87.96	85.60	86.47	86.91
Llama2-13B	71.63	79.91	81.00	76.68	77.81	88.26	88.91	85.83	86.99	87.50
Llama3-8B	73.89	81.19	81.03	81.16	79.32	88.52	88.61	86.45	87.03	87.65
Mistral-7B	76.18	78.46	80.04	76.64	77.83	87.73	88.05	85.72	86.11	86.90
Qwen2.5-7B	85.40	87.35	86.15	86.56	86.37	88.89	89.24	86.78	88.17	88.27
Llama-3.1-8B	80.19	85.73	84.94	84.90	83.94	88.89	88.32	85.64	86.72	87.39
TowerBase-7B-v0.1	85.43	88.01	86.10	88.44	86.99	88.13	88.41	85.16	87.74	87.36
GPT-3.5	87.53	86.97	86.40	86.46	86.84	89.14	89.42	87.47	88.41	88.61
GPT4	88.18	88.13	86.64	85.33	87.07	89.71	89.56	87.41	88.14	88.71
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	89.39	89.22	87.23	89.37	88.80	90.01	89.92	87.55	88.54	89.01

E Additional experimental results

E.1 Crosslingual capabilities of LLMs

Machine translation Tb. 6 report the COMET score on FLoRes-101 benchmark (Goyal et al., 2022) for two directions per language: en → X and X → en. It shows that multilingual LLMs;s translation ability is quite competitive when compared to translation models explicitly trained on parallel corpora or industrial-grade translation APIs.

Embedding analysis Here verify whether the embeddings for a given text in English are similar to the embeddings when some words are presented in different languages. The embedding is a vector as the average of the last layer's activations across all tokens in the sentence (BehnamGhader et al., 2024).

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

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

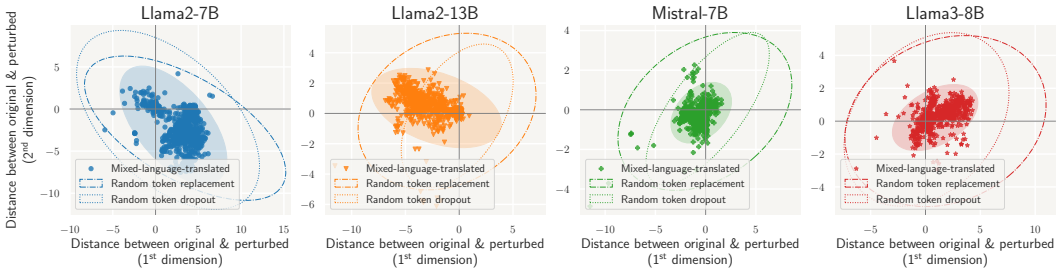


Figure 12: The embeddings of the original English text and the mixed-language-translated text are closely aligned, unlike baselines with unrelated perturbations (e.g., random token replacement or dropout). The ellipses represent the covariance confidence intervals.

Setup. We randomly sample $1k$ examples from the WikiText-103 corpus (Merity et al., 2017), creating two **versions** for each: (1) The original text in **English**; (2) **Mixed-language-translated**: for each word, with a probability of 0.2 it is (independently) translated into a random language selected from {`en`, `fr`, `de`, `es`, `it`} using Google Translate API. That is, each word has a $p = 0.16$ probability of being translated into a non-English language. 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 space of the tokenizer; and (2) **Random Token Dropout**: with a probability of $p = 0.16$, a token is completely masked out by disallowing any attention to it.

To visualize and compare embeddings, we reduce the original 4096-dimensional vectors to 2D using non-linear dimensionality reduction. We then calculate and visualize the per-coordinate distances between 2D embeddings of original English text and mixed-language translations, comparing these to baseline scenarios.

As shown in Fig. 12, for the four LLMs multilingual, including Llama2-7B, Llama2-13B, Mistral-7B and Llama3-8B[9], the embeddings of the original text and its mixed-translated counterpart exhibit a high degree of similarity, with their difference vectors clustering around the origin. This observation stands in stark contrast to the scenario where English words are replaced with random tokens. It implies the explicit crosslingual capabilities of the multilingual LLMs.

E.2 Crosslingual knowledge barriers in LLMs

Evaluation strategy We follow prior work on LLM evaluation (Zheng et al., 2024) to use two evaluation strategies: **(1)** Open-source: access the output probabilities of option ID tokens A/B/C/D and predict argmax. **(2)** Closed-source: compare the golden answer with the 1st generated token (decoding temperature=0), as the logits are not available for most closed-source models.

we evaluated the open-source LLMs using the 1st token produced as the answer in Tb. 7. The two evaluation strategies have a minimal impact on accuracy for 5-shot settings, as demonstrations help regularize the output format. The difference is more evident in the 0-shot setting, likely related to the specific tokenizers. E.g., Llama3-8B treats the 2 characters " A" as 1 token, and has a slightly higher accuracy when using the 1st generated token as the answer. Conversely, Llama2-7B and Mistral-7B tokenizers treat "A" as 1 token. Using the option ID token with the highest probability as the answer for those models generally leads to higher accuracy because it disregards the generation probability of other irrelevant tokens, e.g., "\n", " ".

Crosslingual evaluation of additional models on MMLU knowledge We present additional results for Llama2-7B, Zamba-7B, and Mixtral-8x7B. Figure 13 shows the monolingual

[9]We focus primarily on open-source models due to the cost associated with querying embeddings from proprietary models.

Table 7: Comparing two evaluation strategies for open-source LLMs: option ID token with maximum probability and first new token.

Model	Eval	English MMLU		Mixup MMLU	
		Max prob	1st token	Max prob	1st token
Llama2-7B	0-shot	**41.53**	37.74	**32.18**	27.47
	5-shot	45.88	**45.90**	36.92	**36.96**
Mistral-7B	0-shot	**60.21**	58.41	**47.86**	42.29
	5-shot	**62.57**	62.54	**51.07**	51.05
Llama3-8B	0-shot	60.54	**62.11**	48.62	**50.13**
	5-shot	65.00	**65.39**	51.65	**52.01**

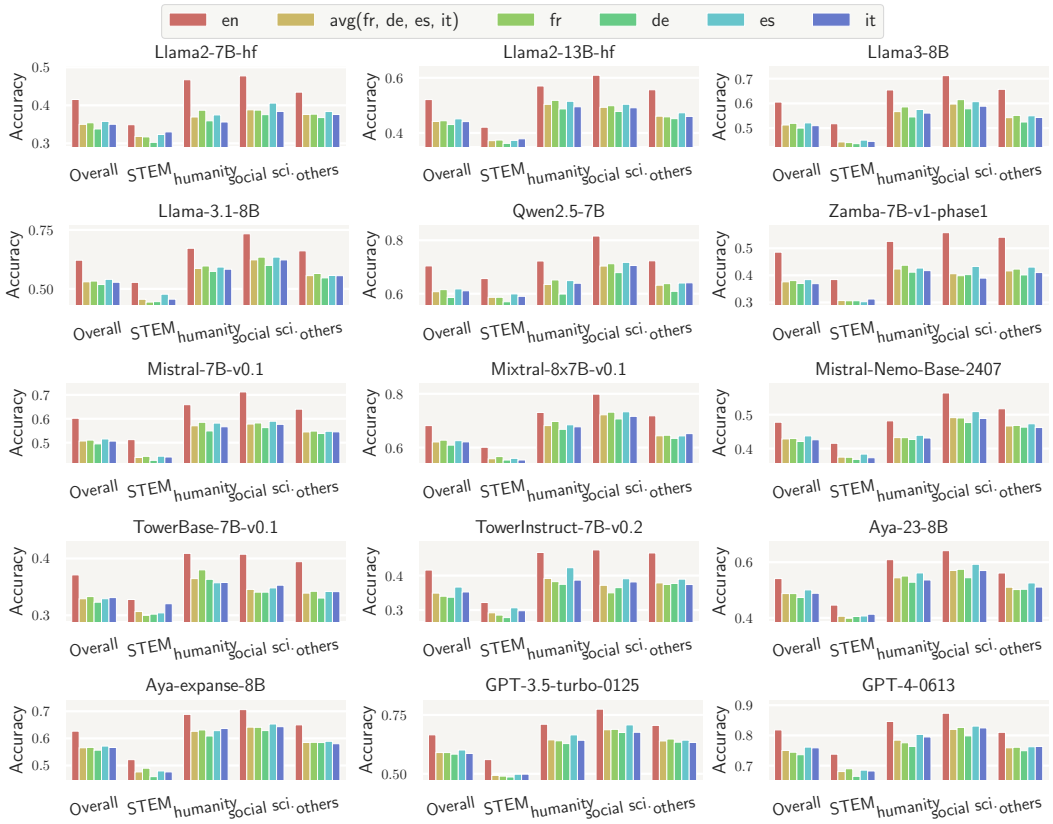


Figure 13: Monolingual evaluation of LLMs on MMLU (fully translated for non-English languages). LLMs consistently perform better at answering multi-choice questions in English than in other languages.

evaluation of LLMs on MMLU, fully translated for non-English languages. The models consistently achieve higher accuracy in English compared to other languages.

Figure 14 displays the accuracy of LLMs on MMLU variant benchmarks. We observe a significant drop in accuracy under crosslingual MCQ evaluation, especially for ground-truth translated MMLU variant, indicating cross-lingual knowledge barriers. The barrier is more pronounced in Llama2-7B and Zamba-7B than in Mixtral-8x7B, possibly due to the larger capacity and multilingual capabilities of Mixtral-8x7B.

E.3 Mixed-language fine-tuning

HP Quiz evaluation results of LLMs fine-tuned on WikiText-2 We finetune for one epoch, with a learning rate of 2×10^{-5} and a batch size of 32. Fig. 8 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). Fig. 15 presents additional results for the Llama2-13B (left) and Mistral 7B (right) models. (1) The trends are consistent with

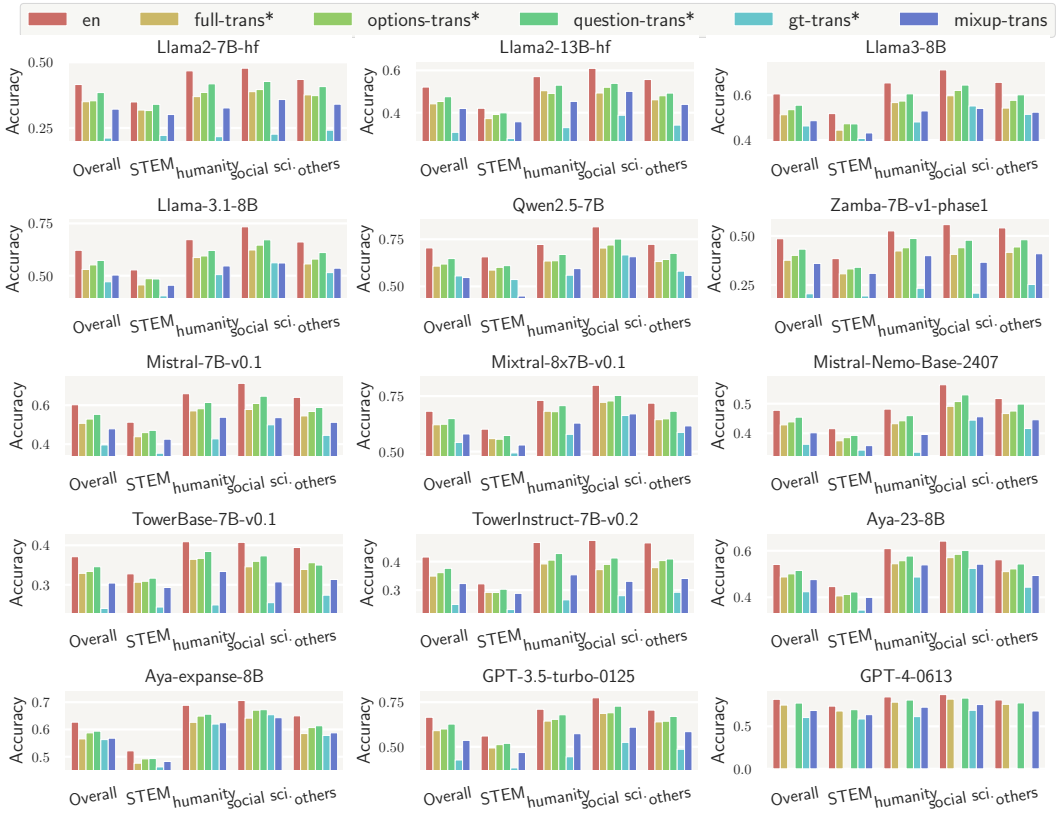


Figure 14: Crosslingual evaluation of LLMs on MMLU variant benchmarks. The bars with * denotes the average accuracy across {`fr`, `de`, `es`, and `it`}. LLMs perform better at answering MCQs in English than in mixed-language settings, especially the ground truth option and mixup translation, indicating the existence of cross-lingual knowledge barriers. Due to budget constraints, GPT-4 is evaluated only in the most challenging settings.

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.

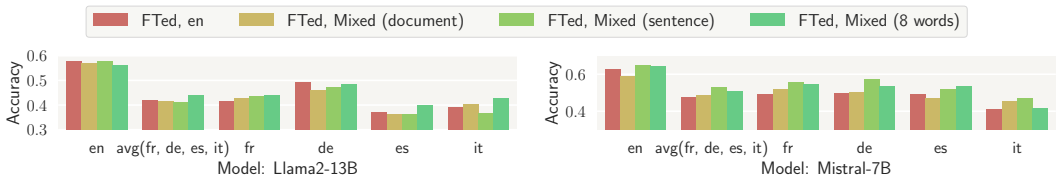


Figure 15: Fine-tuning on a mixed-language general corpus (e.g., WikiText-2) enhances the model's performance on domain-specific task (e.g., Harry Potter knowledge test) across multiple languages, including English.

MMLU evaluation results of LLMs fine-tuned on WikiText-103 Tb. 2 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. 16) and Llama3-8B (Fig. 17) 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. 16 and Fig. 17, 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 non-fine-tuned models, especially in the GT-option translated and mixup translated MMLU setups. These two evaluation setups originally had the lowest performance for the non-fine-tuned 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 non-fine-tuned 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.

Figure 16: 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 non-fine-tuned model.

Figure 17: 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 non-fine-tuned model.

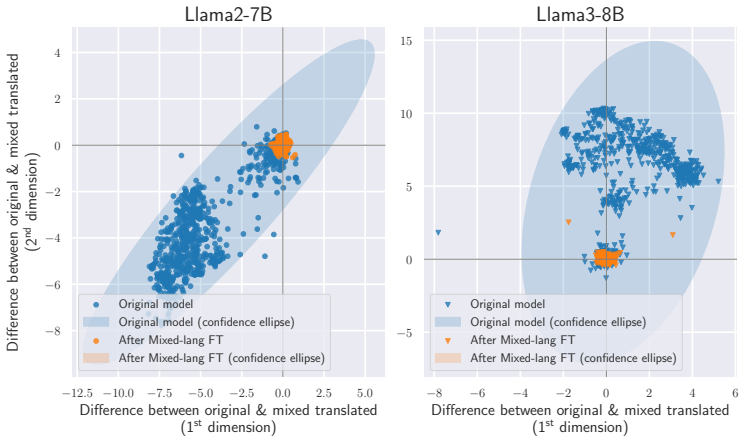


Figure 18: After sentence-level mixed-lang FT, embeddings of original English text & mixed-language-translated text are more closely aligned, indicating a stronger knowledge correlation between En & other langs.

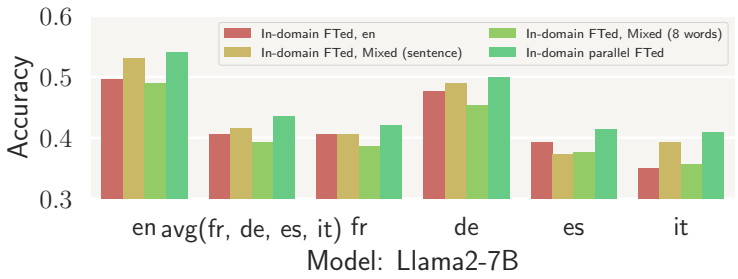


Figure 19: Fine-tuning on a mixed-language domain-specific corpus (i.e., Harry Potter related documents from WikiText-103) generally enhances the performance on the Harry Potter Quiz dataset across multiple languages, including English.