Selected Languages are All You Need for Cross-lingual Truthfulness Transfer

Weihao Liu[*], **Ning Wu**, **Wenbiao Ding**,
Shining Liang, **Ming Gong**, **Dongmei Zhang**
Microsoft STC Asia, Beijing
liuweihao2022@outlook.com
{wuning, wenbiaoding, shiningliang, migon, dongmeiz}@microsoft.com

Abstract

Truthfulness stands out as an essential challenge for Large Language Models (LLMs). Although many works have developed various ways for truthfulness enhancement, they seldom focus on truthfulness in multilingual scenarios. Meanwhile, contemporary multilingual aligning technologies struggle to balance numerous languages and often exhibit serious truthfulness gaps across different languages, especially those that differ greatly from English. In our work, we extend truthfulness evaluation to multilingual contexts and propose a practical method for cross-lingual truthfulness transfer called *Fact-aware Multilingual Selective Synergy* (FaMSS). FaMSS is able to select an optimal subset of all tested languages by language bias and transfer contributions, and then employ translation instruction tuning for cross-lingual truthfulness transfer. Experimental results demonstrate that our approach can effectively reduce the multilingual representation disparity and boost cross-lingual truthfulness transfer of LLMs. [1]

1 Introduction

Large language models (LLMs) have strong ability in generating human-level text in many domains (Brown et al., 2020; Jiang et al., 2023; Touvron et al., 2023). However, almost all LLMs face the issue of generating hallucinated responses (Zhang et al., 2023b), and demand methods to enhance their truthfulness. Although many methods for truthfulness enhancement in English have been proposed (Tonmoy et al., 2024; Wang et al., 2023b), including designing new decoding strategies (Shi et al., 2023; Chuang et al., 2023), synthesizing high quality training data (Tian et al., 2023) and representation editing (Zhang et al., 2024), there is little attention to truthfulness in
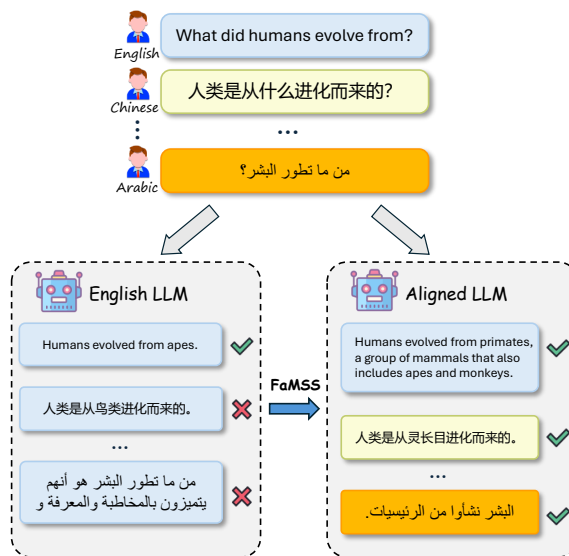

Figure 1: FaMSS is able to align the multilingual capabilities of LLMs, making them more truthful in answering multilingual questions.

multilingual scenarios. Contemporary LLMs generally strive to achieve powerful multilingual capabilities, which promotes their applications worldwide (Team et al., 2023; AI@Meta, 2024). There, exploring the evaluation of truthfulness in multilingual contexts and cross-lingual truthfulness transfer holds significant importance.

To fill this gap of multilingual truthfulness, we construct a benchmark tailored for truthfulness evaluation in multilingual scenarios. Furthermore, we make preliminary explorations into methods aimed at cross-lingual truthfulness transfer of LLMs, with a focus on aligning the internal representations of factual descriptions. The most commonly used technique for building multilingual large models (MLLMs) is cross-lingual instruction tuning (Kulshreshtha et al., 2020; Muennighoff et al., 2023), which usually requires synthesizing multilingual parallel corpus for various down-streaming tasks (Taori et al., 2023; Li et al.,

---

[*]Work done during an internship at Microsoft.
[1]Code is available at https://github.com/microsoft/FaMSS

2024a) and struggles to overcome the alignment difficulties caused by introducing too many languages. In this work, we propose *Fact-aware Multilingual Selective Synergy* (FaMSS), a practical approach for multilingual alignment that relies solely on task-agnostic bilingual data. Meanwhile, FaMSS can achieve significant cross-lingual truthfulness transfer with selected optimal subset of tested languages by language bias and transfer contributions. As illustrated in Figure 1, models utilizing FaMSS demonstrate higher truthfulness in multilingual question-answering tasks.

In summary, our contributions are as follows:

- We construct **MTruthfulQA**, a novel benchmark designed to evaluate the truthfulness of LLMs in multilingual scenarios. This benchmark encompasses nine languages, each containing the same set of questions to ensure equitable evaluation of multilingual capabilities.

- We introduce a practical method for cross-lingual truthfulness transfer called FaMSS. Through a data selection process prior to fine-tuning, FaMSS can efficiently and significantly boost truthfulness of LLMs across multiple languages.

- We systematically investigate how FaMSS facilitates the transfer of truthfulness across multiple languages. Based on our findings, we conclude that simply mixing training data from many different languages, which may interfere with each other, is not the most effective approach.

2 Related Work

2.1 Truthfulness Evaluation Benchmark

For question-answering tasks, TruthfulQA (Lin et al., 2022) provides 817 challenging questions including 38 topics. FreshQA (Vu et al., 2023) is a dataset aimed at testing LLMs on up-to-date world knowledge. For long-form text generation, FActScore (Min et al., 2023) evaluates the factuality of LLMs by breaking down long texts into atomic claims. However, these classic benchmarks for truthfulness evaluation are all in English. In our work, we take a step towards evaluating and enhancing the truthfulness of LLMs in a multilingual setting.

2.2 Multilingual Ability Evaluation

Previous works have provided several datasets and metrics for multilingual fact verification (Gupta and Srikumar, 2021; Qi et al., 2023), multilingual summarization (Aharoni et al., 2022; Qiu et al., 2023) and open-book cross-lingual question-answering (Zhang et al., 2021; Lewis et al., 2020) tasks. With the development of LLMs, many more flexible and comprehensive evaluation datasets have been released. Meanwhile, many evaluation tasks have been expanded to multilingual settings. For instance, MGSM (Shi et al., 2022) evolved from GSM8K (Cobbe et al., 2021) provides mathematical ability evaluation in multilingual contexts. SeaEval (Wang et al., 2023a) hand-crafted high-quality Cross-MMLU and Cross-LogiQA datasets to test general knowledge question answering and reasoning capabilities of LLMs in multilingual scenarios. In addition to these, numerous studies have employed translation tools to construct relevant datasets or devise various metrics for MLLMs (Lai et al., 2023; Lin et al., 2024; Chen et al., 2023; Shafayat et al., 2024; Li et al., 2024b) .

2.3 Handling Numerous Languages

Most works leverage cross-lingual pretraining and fine-tuning on different tasks to achieve multilingual capabilities in their models (Qin et al., 2024; Lample and Conneau, 2019; Huang et al., 2019; Devlin et al., 2019; Xue et al., 2021; Muennighoff et al., 2023; Zhu et al., 2023). However, the difficulty in learning multiple languages increases significantly with the number of languages. Conneau et al. (2020) discussed the challenges of scaling MLLMs to more languages, such as the *curse of multilinguality*, where increasing the number of languages can dilute the model's capacity and potentially decrease overall performance. Therefore, some works mainly focus on training with high-resource languages(Yang et al., 2022) or mining specific language pairs (Fan et al., 2021a; Lin et al., 2023). Other approaches propose clustering target languages into groups and training a single model for each cluster (Tan et al., 2019; Fan et al., 2021b).

3 Multilingual Truthfulness Benchmark

3.1 Data Source

Since we mainly focus on the truthfulness (Sun et al., 2024) of LLMs rather than their faithfulness (Es et al., 2023; Maynez et al., 2020), we prefer to consider context-free evaluation tasks. We

employ TruthfulQA (Lin et al., 2022), a widely used dataset for evaluating the truthfulness of models in a question-answering format, as our foundation data. Compared with other QA datasets, TruthfulQA places greater emphasis on the ability to avoid false beliefs or misconceptions rather than complex reasoning.

3.2 Multilingualism Formulation

To build a multilingual benchmark, we first select a few target languages from a diverse linguistic pool. Our evaluation system incorporates nine different languages: *English (En)*, *French (Fr)*, *German (De)*, *Spanish (Es)*, *Chinese (Zh)*, *Japanese (Ja)*, *Russian (Ru)*, *Thai (Th)* and *Arabic (Ar)*. For the questions and answers in original dataset, we translate them into the target languages through prompting *GPT-4* [2]. The prompt template we used is shown in Appendix A.1.

Considering the instability of LLM-based translation, we also recruited native speakers to validate the translations. Each translation is reviewed by three native speakers, who vote on its correctness. If at least two of the three agree, the translation is considered accurate. The average accuracy of the translation is 96.4% including both the questions and corresponding answers. We further manually checked and rectified these translation errors.

3.3 Evaluation Metrics Migration

We migrate and improve the evaluation metrics used in TruthfulQA (Lin et al., 2022). For multi-choice QA, we adopt three likelihood scores $MC1$, $MC2$ and $MC3$, which is similar to Chuang et al. (2023). Detailed calculations of metrics are reported in Appendix A.2.

For open-ended generation, we replace the evaluation model in the original pipeline with a stronger open source model Mistral-7b [3] (Jiang et al., 2023). We then utilize a similar translation pipeline to build the multilingual classification dataset and fine-tune two models for truthfulness and informativeness evaluation, which we refer to as "MM-Judge (Multilingual-Mistral-Judge)". An acceptable answer should be both truthful and informative (e.g. "I have no comment." is truthful but lacks valuable information). These two scores to judge whether the answer is truthful and informative are denoted as True (%) and Info (%) respectively. The

product True*Info (%) serves as a more comprehensive and reasonable metric. The reliability of "MM-Judge" is discussed and verified in Appendix E.

4 Fact-aware Multilingual Selective Synergy

To balance and enhance truthfulness in different languages, we propose Fact-aware Multilingual Selective Synergy (FaMSS), which can effectively fine-tune models for multilingualism with fact-aware multilingual data in a selective languages synergy manner.

4.1 Fact-aware Multilingual Data

Our primary training data is derived from a variety of parallel corpus. Although there are many kinds of parallel corpora used in machine translation, including WikiMatrix (Schwenk et al., 2021), UNPC (Ziemski et al., 2016), Tatoeba (Tiedemann, 2020), etc., we consider not using too much common machine translation corpora. These MT corpora mainly consist of simple sentences in each language with limited factual descriptions or only focus on specific areas, which does not contribute much to improving the truthfulness of models. Consequently, we employ fact-aware multilingual data, which includes rich bilingual factual descriptions alongside high-quality parallel corpora. In general, we include three different types of data:

Factuality Translation Data Similar to MLQA (Lewis et al., 2020), we first perform parallel sentence mining over Wikipedia articles through the LASER toolkit[4] (Artetxe and Schwenk, 2019) with some specific topics, including *history*, *biography*, *geography*, *science* and *cultural*. For each parallel sentence, we extract a broader context (e.g. a paragraph) that provides additional factual details related to the sentence. This encourages the factual integrity and depth of the content. Given the challenges of exploring fully aligned context across nine languages, we also maintain a 4-way aligned approach, which means that a factual description appears in at least four different languages (always including English).

Common Translation Data In our study, we find that a certain amount of common translation data is still beneficial for multilingual alignment. FLORES-200 (Costa-jussà et al., 2022) provides high-quality translation data between more than

---

200 languages, making it a suitable choice for our purposes.

Pretraining Data   Dou et al. (2024) find that excessive amount of instruction data during supervised fine-tuning (SFT) can degrade the world knowledge stored in LLMs, leading to performance drop in knowledge tasks. Our experiments also support this conclusion. Thus, besides translation data, we incorporate extra English pretraining data from Wikipedia to alleviate the world knowledge forgetting. Note that we usually have about 10% of the training data as pretraining data.

We have data of all nine languages mentioned in section 3, which means a massive number from possible translation directions ($9 \times 8 = 72$) for a single example. Therefore, similar to Zhu et al. (2023), we follow the most widely used English-centric way and further adopt the best translation direction, which means that we only put the non-English text on the target side and the English text on the source side. The details of our collected training data are presented in Appendix B.2.

4.2   Selective Languages Synergy

Incorporating a large number of languages complicates the alignment process, which is known as the *curse of multilinguality* (Conneau et al., 2020). Therefore, we aim to select a core subset of languages, which helps align all nine languages in a collaborative and diffusive manner. This idea also coincides with Pareto principle (Sanders, 1987), which indicates that the most important factors in a system are only about 20% but affect 80% of the results. In the selection process, we first propose a **language bias probe** to help cluster languages into several groups and then select core languages from different groups.

4.2.1   Language Bias Probe

Different languages have different characteristics, and some of them may share representation similarities or show big differences, which can be used to guide the selection of training data. We propose a language bias probe to estimate the representation bias of models between languages. Inspired by the findings of Ju et al. (2024) that different layers in LLMs may encode different contextual information, we hypothesize that models exhibit minimal mean language bias in the layer where they encode richest semantic representations. For convenience, we will refer to the layer encoding
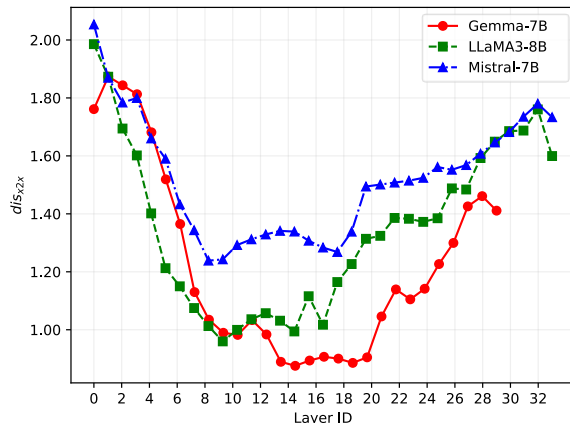

Figure 2: Mean bias between languages. dis_{x2x} represents the average distance value of all language pairs in one layer.

the most language-agnostic semantic information as the semantic layer.

Formally, we define the bias between each language pair as their average representation distance on the probe corpora $\mathcal{C}$, denoted as $\mathcal{D}$. The algorithm to calculate $\mathcal{D}$ is described in Algorithm 1, where $\mathcal{C}$ comes from FLORES-200 (Costa-jussà et al., 2022) in our work. As shown in Figure 2, the mean bias of languages in three popular foundation models initially decreases and then increases as the index of decoder layers increases. Thus, we can infer that the lowest point on each curve indicates the semantic layer of each foundation model (e.g., layer 14 may be the semantic layer of *Gemma-7B*). We will leverage this information to analyze the transfer contributions of specific languages and optimize the combination of training data.

4.2.2   Probing to Select

Suppose that we have M languages in our training corpora $\mathcal{C}$ and evaluation set denoted as $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$. Our goal is to select an optimal language set s_o from $\mathcal{L}$, which can boost the alignment between English and other non-English languages as effectively as possible.

We take two major steps to form s_o: 1) group all languages into several clusters and 2) select a core language from each cluster as one element in s_o. For step 1, we cluster languages by language bias. For step 2, we define a transfer contribution $\mathcal{TC}$ for each language:

$$\mathcal{TC}_l = \sum_{l' \in \mathcal{L}} \mathcal{D}_s[\text{"English"}][l'] - \mathcal{D}_s^l[\text{"English"}][l']$$

where $\mathcal{D}_s, \mathcal{D}_s^l$ means the distance matrix of the

Algorithm 1 Language Bias Probe

Input: Model $\mathcal{M}$, parallel corpus $\mathcal{C}$, language set $\mathcal{L}$.

Output: Bias $\mathcal{D}$ between any language pair in $\mathcal{L}$ of each layer in $\mathcal{M}$.

1: $\mathcal{L} \leftarrow \{$"English", "Chinese", "Spanish", ... $\}$
2: $N \leftarrow$ number of decoder layers in $\mathcal{M}$
3: $M \leftarrow$ number of samples in $\mathcal{C}$
4: **for** $i = 0$ **to** N **do**
5:    $\mathcal{H}[i] \leftarrow$ mean sentence hidden states calculated from $\mathcal{C}$ at layer i
6:    $\mathcal{H}[i] \leftarrow$ Standardization($\mathcal{H}[i]$)
7:    **for** (l_j, l_k) **in** $(\mathcal{L}, \mathcal{L})$ **do**
8:      **for** $m = 0$ **to** $M - 1$ **do**
9:        $d \leftarrow \|(\mathcal{H}[i][l_j][m] - \mathcal{H}[i][l_k][m])\|_2^2$
10:        $\mathcal{D}[i][l_j][k] \leftarrow \mathcal{D}[i][l_j][k] + d$
11:      **end for**
12:      $\mathcal{D}[i][l_j][l_k] \leftarrow \mathcal{D}[i][l_j][l_k]/M$
13:    **end for**
14: **end for**
15: Return $\mathcal{D}$

Algorithm 2 Optimal Language Set Selection

Input: Language set $\mathcal{L}$, language bias matrix $\mathcal{D}_s$ of the semantic layer, transfer contribution of each language $\mathcal{TC}$, maximum number of languages m, distance threshold d.

Output: Optimal language set s_o

1: $\mathcal{S} \leftarrow \{\, \{l\} \mid l \in \mathcal{L} \,\}$
2: **while** $|\mathcal{S}| > m$ **do**
3:    $\mathcal{S}' \leftarrow \mathcal{S}$
4:    **for** each $s \in \mathcal{S}$ **do**
5:      $s' \leftarrow$ nearest_set($s, \mathcal{S}, \mathcal{D}_s, d$)
6:      **if** s' exists **then**
7:        $\mathcal{S}' \leftarrow (\mathcal{S}' \backslash \{s, s'\}) \cup \{s \cup s'\}$
8:      **end if**
9:    **end for**
10:    **if** $|\mathcal{S}| = |\mathcal{S}'|$ **then**
11:      break
12:    **end if**
13:    $S \leftarrow S'$
14: **end while**
15: $s_o \leftarrow \{$element_with_max_$\mathcal{TC}(s) \mid s \in \mathcal{S}\}$
16: Return s_o

---

semantic layer before and after fine-tuning on language l. The language with largest $\mathcal{TC}$ in corresponding cluster is selected as the core language of that cluster. Algorithm 2 illustrates the algorithm to select optimal language set s_o, where the role of the parameters m and d is to control the number of languages involved in the training stage and to avoid merging languages with too large differences into the same set. For the nearest set, we define the distance between two sets s and s' as the minimum language bias among all language pairs (l_1, l_2) in $s \times s'$. For more details about the selecting process, please refer to Appendix C.

4.2.3 Translation Instruction Tuning

We adopt the instruction tuning (Zhang et al., 2023a) pipeline leveraging constructed bilingual data. We use the translation instruction template in Figure 8 of Appendix B.3, and fine-tuning models by next-token prediction with cross-entropy loss.

5 Experiments

5.1 Setup

In our experiments, we utilize three representative open source foundation LLMs as our base models: *LLaMA-3-8B* (AI@Meta, 2024), *Mistral-7B-v0.1* (Jiang et al., 2023) and *Gemma-7B* (Team et al., 2024). We employ full parameters fine-tuning over selected subset of fact-aware multi-

lingual data. Details about our training are placed in Appendix B.

For evaluation benchmarks, we first evaluate our models on MTruthfulQA we construct in both multi-choice QA and open-ended generation settings. We also do evaluations on Cross-MMLU (Wang et al., 2023a), which contains 150 high-quality human-annotated questions for common knowledge testing across seven languages: *English*, *Chinese*, *Indonesian*, *Spanish*, *Vietnamese*, *Malay*, and *Filipino*.

5.2 Main Results

Table 1 and Table 2 present the results of models on MTruthfulQA. Following the standard setting from Lin et al. (2022) and Zhang et al. (2024), all experiments are performed in a few-shot prompting setting. We analyze the performance of all models in depth and draw some conclusions.

MTruthfulQA reveals the ability of LLMs' truth-telling and truth-recognizing in multilingual scenarios. First, in open-ended generation evaluation, strong baselines such as *GPT-4* (Achiam et al., 2023) and *Mistral-7B-Instruct-v0.2* (Jiang et al., 2023) get much higher scores compared with other models, showing more powerful truth-telling ability across massive languages.

Models	True*Info(%)									
	En	De	Fr	Es	Ru	Zh	Ja	Th	Ar	Avg.
GPT-4 (Achiam et al., 2023)	69.8	71.4	72.5	67.9	68.9	68.9	68.1	56.5	68.5	68.1
Bloomz-7B1-mt (Muennighoff et al., 2023)	18.7	21.9	20.7	13.2	16.0	26.8	16.4	13.5	14.2	17.9
LLaMA-3-8B-Instruct (AI@Meta, 2024)	56.7	30.7	43.7	29.1	37.3	34.4	18.7	17.3	31.6	33.3
Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	66.7	52.5	57.9	49.9	49.1	52.9	27.7	24.6	42.7	47.1
With FaMSS										
LLaMA-3-8B (AI@Meta, 2024)	36.6	31.6	31.1	27.3	38.8	31.1	25.1	20.1	42.1	31.5
LLaMA-3-8B + FaMSS	**44.9**	**34.9**	**34.5**	**37.3**	**42.4**	**36.2**	**39.9**	**24.7**	41.9	**37.4**
Mistral-7B-v0.1 (Jiang et al., 2023)	36.0	35.9	28.8	29.3	34.0	35.9	19.5	15.4	22.2	28.5
Mistral-7B-v0.1 + FaMSS	**39.6**	**41.2**	**31.4**	29.3	33.0	**40.0**	**20.5**	14.5	**29.6**	**31.0**
Gemma-7B (Team et al., 2024)	25.7	19.7	17.7	17.1	21.5	27.8	34.6	15.8	19.7	22.2
Gemma-7B + FaMSS	**37.8**	**45.9**	**22.2**	**23.6**	**30.8**	**41.7**	**40.1**	**19.7**	**25.2**	**31.9**

Table 1: Experimental results on MTruthfulQA dataset under open-ended generation setting. Presented metric is True*Info (%) score, representing ability to generate truthful and informative contents. Full results can be found in Appendix D.

Models	MC1 (%)									
	En	De	Fr	Es	Ru	Zh	Ja	Th	Ar	Avg.
Bloomz-7B1-mt	26.7	27.1	27.3	24.4	29.0	26.2	29.7	26.2	27.7	27.1
LLaMA-3-8B-Instruct	40.6	36.4	40.5	37.5	35.7	37.9	36.0	28.5	35.7	36.5
Mistral-7B-Instruct-v0.2	55.4	49.7	52.8	47.4	44.7	42.2	40.3	29.7	37.5	44.4
With FaMSS										
LLaMA-3-8B	32.4	31.5	33.5	30.5	31.6	28.9	34.6	28.3	29.9	31.2
LLaMA-3-8B + FaMSS	**35.0**	**31.7**	32.7	**34.5**	**32.8**	**30.2**	**35.5**	26.9	**30.1**	**32.2**
Mistral-7B-v0.1	30.6	33.5	33.0	29.7	33.0	32.1	32.9	28.2	31.3	31.6
Mistral-7B-v0.1 + FaMSS	**32.8**	**33.7**	32.8	**31.6**	**33.4**	**32.9**	32.7	27.7	**31.8**	**32.1**
Gemma-7B	33.4	32.4	33.2	33.4	30.5	29.9	33.0	27.7	31.7	31.7
Gemma-7B + FaMSS	**37.3**	**34.8**	**36.2**	**33.9**	**31.3**	**30.4**	**33.4**	**28.9**	31.6	**33.1**

Table 2: Experimental results on MTruthfulQA dataset under multi-choice QA setting. Presented metrics are $MC1$ scores mentioned in section 3. Full results can be found in Appendix D.

Similar results are observed for truth-recognizing in multi-choice setting. Furthermore, we can obviously see models exhibit different truthfulness levels across different languages, usually higher in English or other languages similar to English. This is reasonable given that LLMs are trained with more data in these languages. Additionally, the gap between different language pairs is not uniform in the two evaluation settings. We can see larger gap in True*Info (%) scores compared with MC scores, indicating that open-ended generation is a more challenging task and better distinguishes the multilingual capabilities of models.

FaMSS is convenient and effective. Although we just train our models with a majority of translation instruction data, our method achieves great enhancement of truthfulness in tested foundation models. We get **+5.9%** of improvement in True*Info (%) score on *LLaMA-3-8B*, **+2.5%** on *Mistral-7B-v0.1* and **+9.7%** on *Gemma-7B*. Meanwhile, the result of multi-choice QA on MTruth-

fulQA is also averagely positive. Additionally, models fine-tuned with FaMSS show stronger multilingual capabilities in common knowledge benchmark. As shown in Table 3, our method is significantly effective on Cross-MMLU, raising accuracy of prediction by **+5.2%** on *LLaMA-3-8B*, **+1.4%** on *Mistral-7B-v0.1* and **+5.6%** on *Gemma-7B*. A notable phenomenon is that the performance improvement of *Mistral-7B-v0.1* is relatively small. This may be due to the relatively limited vocabulary size of it compared with other models, which restricts the efficiency of multilingual alignment and highlights the importance of a large vocabulary for MLLMs. In general, these positive results highlight the effectiveness of FaMSS, showing that fact-aware multilingual data can be conveniently used to enhance multilingual truthfulness.

Trained and not directly trained languages can both benefit from FaMSS. We investigate the difference in performance change between trained languages and untrained languages. For exam-
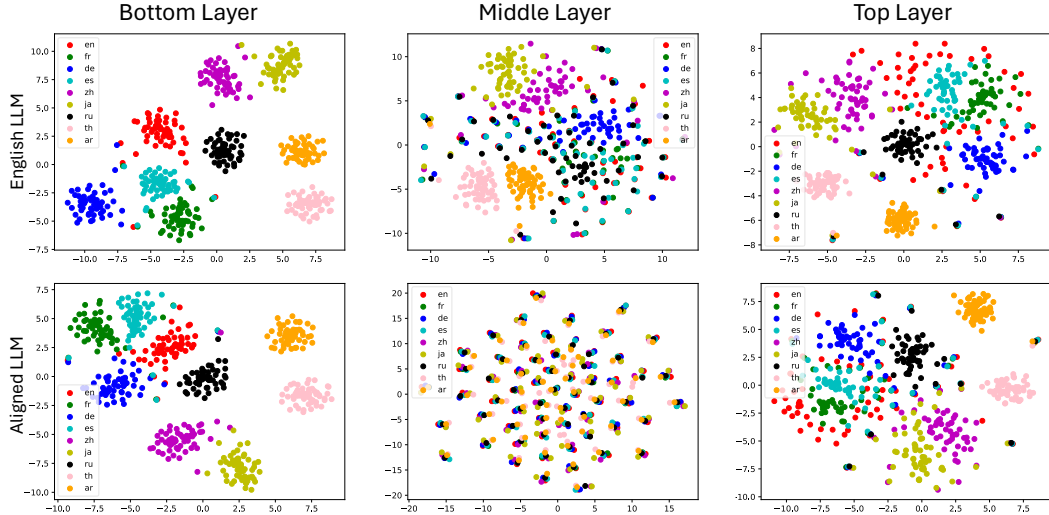
Figure 3: Visualization results in the representation space of LLMs on FLORES-200 before and after utilizing FaMSS. Our method successfully enhances the multilingual representation alignment in the middle layers, where models encode more semantic-level information. In contrast, the top and bottom layers contain more syntactic information, resulting in little overlap in representations across different languages.

Model	Acc	Consis	AC3
GPT-4	85.0	80.0	83.0
LLaMA3-8B-Instruct	57.0	48.0	52.0
Mistral-7B-Instruct-v0.2	49.0	30.0	37.0
With FaMSS			
LLaMA-3-8B	46.8	28.6	35.5
LLaMA-3-8B + FaMSS	**52.0**	**40.8**	**45.7**
Mistral-7B-v0.1	51.2	45.1	48.0
Mistral-7B-v0.1 + FaMSS	**52.6**	**46.3**	**49.2**
Gemma-7B	48.4	54.2	51.1
Gemma-7B + FaMSS	**54.0**	52.0	**53.0**

Table 3: Experimental results on Cross-MMLU dataset.

ple, when we fine-tuned *Gemma-7B* on language set $\mathcal{S}$ = {German, Chinese, Arabic}, our tuning method get higher True*Info (%) scores in these languages in $\mathcal{S}$ compared to the base model, which can be regarded as these languages "learn" from English during FaMSS. On the other hand, languages not in $\mathcal{S}$ also benefit from the alignment procedure between English and $\mathcal{S}$, showing relatively smaller improvement on True*Info (%) score. Moreover, the performance gain on Cross-MMLU further support this conclusion, as there are very few languages that are both included in Cross-MMLU and $\mathcal{S}$ (i.e., just Chinese since English is the source language). To get a deeper understanding of the enhancement in semantic alignment between languages brought by FaMSS, we also present a visual comparison through the t-SNE [5] algorithm in Fig-

ure 3.

6 Ablation and Further Analysis

To further explore the influence of FaMSS across languages and explain the data allocation schema used in our experiments, we conduct ablation studies and additional analyses. The reported results are obtained on *Gemma-7B*, but similar conclusions were found across other models as well.

6.1 Data Allocation of Languages

Alignment between multiple languages may not be performed synchronously. Ideally, we would combine data from all languages and hope for an aligning process between each individual language and English. However, the representation of English text in semantic space cannot simultaneously align with all other languages. In fact, languages do interfere with each other when trained together. To demonstrate this point, we separately fine-tuned the base model with data in each language and leveraged the language bias probe to measure the movement of language bias between English and other languages. As shown in Figure 4, the representation of a given language X consistently becomes closer to English when fine-tuned on translation data in language X. However, for some languages, this can also cause a larger gap between English and another language Y. For instance, fine-tuning on translating English to French makes French closer to English but negatively impacts the
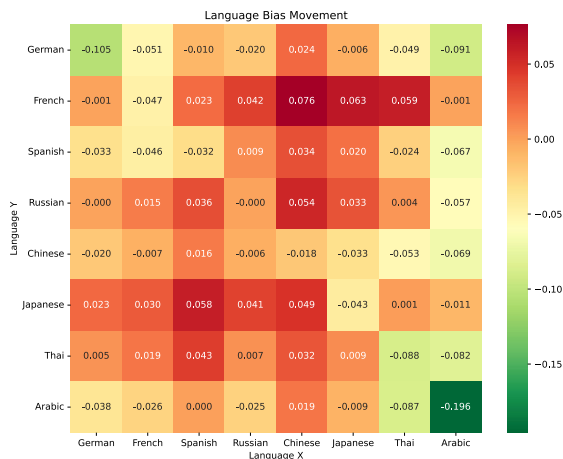
alignment between English and Chinese.

Figure 4: Language bias movement when fine-tuned on language Y and probed on language X. All displayed values are calculated on the 14th layer of *Gemma-7B*. The value can also be formulated as $\mathcal{D}_s^Y[\text{"English"}][X] - \mathcal{D}_s[\text{"English"}][X]$.
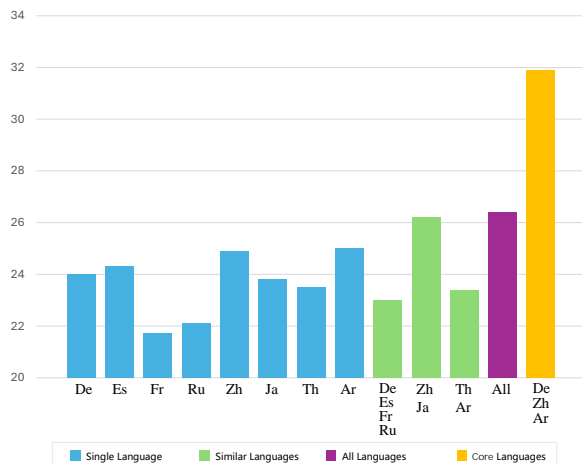

Figure 5: Performance on MTruthfulQA when fine-tuned with different language sets.

Combination of a few languages is more robust. One main advantage of our leveraged FaMSS method is that we can fine-tune models with just a subset of target languages while achieving more robust performance. To further confirm the rationality behind selective languages synergy, we compared the model's performance under different language selection settings. We report the True*Info (%) scores of training with different language sets on *Gemma-7B* in Figure 5. The results show that simply mixing training data from similar languages may actually harm the model's performance, even if the total amount of training data

increases. This indicates that training with data from different languages makes it more challenging for the model to improve its overall multilingual truthfulness. In contrast, selecting only three core languages instead of the whole language set enables the model to learn the intrinsic connections between multiple languages more efficiently, thereby achieving a higher level of multilingual proficiency. We also conducted ablation studies on the performance change when the language set grew larger with different m and d values defined in section 4.2.2, and the results in Table 11 in Appendix D also confirmed that increasing the number of languages could potentially impair the overall multilingual truthfulness.

6.2 Effectiveness of Mixture Data Types

Figure 6 displays the results on *Gemma-7B* with different mixtures of data types. The optimal performance is achieved with a balanced integration of all three data types. The synergistic effect of a diverse data mixture may arise from the complementary nature of different data types (red line).
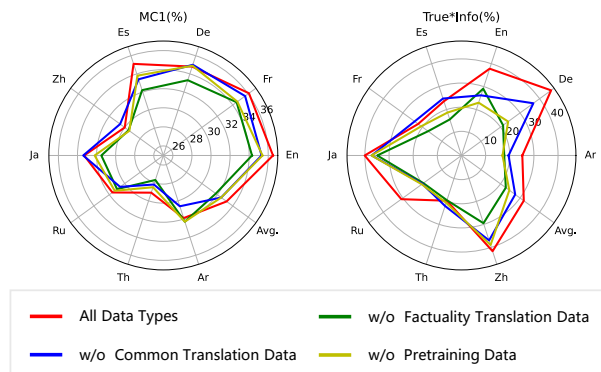
Figure 6: Multilingual performance of LLMs with different mixture of data types.

The omission of factuality data (green line) leads to a noticeable struggle in enhancing the truthfulness of LLMs. This is primarily due to the model's reduced capacity to discern and generate factually accurate information. Meanwhile, the inclusion of the other two types of data further improves the truthfulness level across most languages.

7 Conclusion

In this paper, we construct a novel MTruthfulQA benchmark, which enables the evaluation of truthfulness in multilingual scenarios. Through our experiments, we draw the conclusion that simply mixing data of languages together results in a low

cross-lingual truthfulness transfer efficiency and harm the overall performance. In the contrast, our proposed method FaMSS successfully boosts cross-lingual truthfulness transfer with less well-selected data.

Limitations

In our study, we primarily evaluate our models on closed-book QA tasks. It remains unclear whether our method is equally effective on extractive QA or other context-demanded tasks, which are more related to the faithfulness of LLMs. Regarding the data allocation of different languages, it is possible that incorporating a small proportion of not considered languages rather than entirely omitting them, might yield better results. Considering the numerous scenarios introduced by varying data proportions, we leave it as future research to explore the impact of diverse language ratios on the model's overall multilingual performance.

Ethical Considerations

The translated multilingual data utilized in this work may not be perfectly aligned with the original source data and could contain some unreasonable descriptions. We have made every effort to mitigate these issues to the best of our ability. Additionally, the results in our language selection process might lead to misconceptions regarding the superiority or inferiority of certain languages; however, our work does not contain any bias against any language. Our work strictly adheres to the license and policies of released LLMs and publicly available datasets.

References

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

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. mface: Multilingual summarization with factual consistency evaluation. *arXiv preprint arXiv:2212.10622*.

AI@Meta. 2024. Llama 3 model card.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

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

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Loramoe: Alleviate world knowledge forgetting in large language models via moe-style plugin. *Preprint*, arXiv:2312.09979.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021a. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Yimin Fan, Yaobo Liang, Alexandre Muzio, Hany Hassan Awadalla, Houqiang Li, Ming Zhou, and Nan Duan. 2021b. Discovering representation sprachbund for multilingual pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 881–894.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. *arXiv preprint arXiv:2106.09248*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pretraining with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

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

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching Yun Chang. 2020. Cross-lingual alignment methods for multilingual bert: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. 2024a. Eliciting better multilingual structured reasoning from llms through code. *arXiv preprint arXiv:2403.02567*.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024b. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.

Geyu Lin, Bin Wang, Zhengyuan Liu, and Nancy F Chen. 2024. Crossin: An efficient instruction tuning approach for cross-lingual knowledge alignment. *arXiv preprint arXiv:2404.11932*.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André FT Martins, and Hinrich Schütze. 2023. mplm-sim: Unveiling better cross-lingual similarity and transfer in multilingual pretrained language models. *arXiv preprint arXiv:2305.13684*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. *arXiv preprint arXiv:2305.13632*.

Robert Sanders. 1987. The pareto principle: its use and abuse. *Journal of Services Marketing*, 1(2):37–40.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore. *arXiv preprint arXiv:2402.18045*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

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

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023b. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2022. Hlt-mt: High-resource language-specific training for multilingual neural machine translation. *arXiv preprint arXiv:2207.04906*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023a. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. *arXiv:2108.08787*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A Details of Benchmark Construction

A.1 Prompt for Translation

We use the prompt format in Figure 7 to translate questions and answers in TruthfulQA into other languages.

Figure 7: Prompt to translate questions and answers.

A.2 Metrics for MTruthfulQA

Formally, we define the question in language l as $\mathcal{Q}_l$ and the response of LLM $\mathcal{L}$ as $\mathcal{R}_\mathcal{L}(\mathcal{Q}_l)$. For multi-choice QA, we define $\mathcal{AB}_l$ as the best answer, $\mathcal{AT}_l$ as the set of correct answers and $\mathcal{AF}_l$ as the set of wrong answers in corresponding language l. The three likelihood scores $MC1$, $MC2$ and

$MC3$ are calculated as follows:

$$MC1_l = \begin{cases} 1, & if\ p(\mathcal{AB}_l) > \max_i p(\mathcal{AF}_{l,i}) \\ 0, & else \end{cases}$$

$$MC2_l = \frac{\sum_i \exp(p(\mathcal{AT}_{l,i}))}{\sum_i \exp(p(\mathcal{AB}_{l,i})) + \sum_i \exp(p(\mathcal{AF}_{l,i}))}$$

$$MC3_l = \frac{\sum_i [p(\mathcal{AT}_{l,i}) > p(\mathcal{AF}_{l,i})]}{|\mathcal{AB}_l|}$$

where l represents the language evaluated on, $p(x)$ represents the logits of x and the symbol with subscript i denotes the ith item in the corresponding set. The final MC scores are the average scores of all test samples.

For open-ended generation:

- **True (%)**: the percentage of responses that are classified as truthful.

- **Info (%)**: the percentage of responses that are classified as informative.

- **True*Info (%)**: the percentage of responses that are classified as both truthful and informative.

B Details of the Fact-aware Multilingual Alignment

B.1 Hyperparameters

Hyperparameter	Value
learning_rate	3e-6
batch_size	4
gradient_accumulation_steps	2
epochs	4
model_max_length	2048
lr_scheduler_type	cosine
fp16	True
optimizer	AdamW

Table 4: Fine-tuning hyperparameters.

B.2 Statistics of Training Data

Table 5 shows the statistics on the two types of translation data we collect. Table 6 presents the allocation of training data in our main experiment.

Data Type	De	Fr	Es	Ru	Zh	Ja	Th	Ar
Factuality Translation	4517	4235	5253	5223	5137	4236	4239	5335
Common Translation	997	997	997	997	997	997	997	997

Table 5: Statistics of collected training data in our experiments.

Data Type	Target Language	Number of Items
	De	4517
Factuality Translation	Zh	5137
	Ar	5335
	De	997
Common Translation	Zh	997
	Ar	997
Pretraining	En	1946

Table 6: Detailed information of training data used for FaMSS with optimal allocation of data types and languages.

B.3 Template for Translation Instruction Tuning

> Translation Instruction Tuning Template:
>
> Translate text in language {source language} to {target language}. Remember that they have the same meaning.
> <{source language}>: {source text} => <{target language}>: {target text}

Figure 8: Format of instruction tuning data in translation task.

C Details About Optimal Languages Selection

In the process of optimal language set selection described in Algorithm 2, M is set to 3 and d is set to the average bias between all different language pairs (i.e., $d = 0.84$). As shown in Figure 9, we first merge all languages into three groups, each containing similar languages. Then we calculate the transfer contribution $\mathcal{TC}_l$ according to the mean bias movement for each language. We finally select three core languages from these corresponding three sets (i.e., German, Chinese and Arabic).

D More Results

We report more detailed results evaluated on MTruthfulQA in Table 7, 8, 9 and 10. These metrics include True (%), Info (%), $MC2$ (%) and $MC3$ (%). We also present the ablation studies results over different m, d values in Table 11.

E Reliability of "MM-Judge"

We deploy Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) model as the base model of the evaluator used for open-ended generation. One of the reasons is that this model exhibits the best performance in multi-choice QA task shown in Table 2, which demonstrates that this model is relatively more truthful and strong in distinguishing truthful responses. After fine-tuning the model into two "MM-Judge" models, we also tested their accuracy in judging whether a response is truthful or informative. First, we tested the accuracy of truthful judgement using the correct and incorrect answers provided by MTruthfulQA dataset (we did not test informativeness at this step, since no exact labels of informativeness were provided). The results is listed in Table 12. We also tested the evaluation results of real responses from our main experiments. Specifically, we used all responses generated by **Llama-3-8B+FaMSS**, then we asked a few trained annotators to assess whether the evaluator model's evaluations were correct. The results we obtained are presented in Table 13. The average accuracy of "MM-Judge" is 91.8% in truthfulness and 97.5% in informativeness, which is reasonably robust and reliable.

Models	True (%)									
	En	De	Fr	Es	Ru	Zh	Ja	Th	Ar	Avg.
GPT-4	72.2	72.7	73.7	69.3	70.4	70.4	68.9	57.0	71.5	69.6
Bloomz-7B1-mt	92.5	55.0	92.3	99.5	84.8	76.1	82.1	68.4	91.3	82.5
LLaMA-3-8B-Instruct	71.2	85.7	70.5	77.4	47.7	74.3	91.4	76.5	68.4	76.0
Mistral-7B-Instruct-v0.2	80.0	78.6	76.5	79.2	75.4	69.8	80.2	51.9	69.8	74.1
With FaMSS										
LLaMA-3-8B	58.9	78.6	76.5	79.2	75.4	69.8	80.2	51.9	74.9	74.1
LLaMA-3-8B + FaMSS	60.5	62.5	56.3	65.0	49.4	47.9	57.2	37.9	52.3	54.3
Mistral-7B-v0.1	55.2	59.0	60.8	68.3	61.9	49.7	82.1	67.1	81.5	65.1
Mistral-7B-v0.1 + FaMSS	46.4	59.6	66.7	70.6	63.0	52.6	84.3	64.7	85.9	66.0
Gemma-7B	76.5	85.6	85.2	87.3	80.2	71.2	69.9	76.9	77.7	78.9
Gemma-7B + FaMSS	78.5	84.7	86.0	86.2	73.6	64.7	60.7	68.7	79.2	75.8

Table 7: True (%) score of models evaluated on MTruthfulQA. In our experiment, we find that a single True (%) score does not reflect the actual ability to give truthful answers since models are always refusing to give a clear answer, which also improve their True (%) score.

Models	Info (%)									
	En	De	Fr	Es	Ru	Zh	Ja	Th	Ar	Avg.
GPT-4	97.6	98.4	98.8	98.7	98.4	98.5	99.0	99.1	97.1	98.4
Bloomz-7B1-mt	26.2	61.2	28.4	13.6	30.2	49.8	32.9	36.2	22.8	33.5
LLaMA-3-8B-Instruct	84.9	45.0	72.6	51.7	69.4	60.0	27.2	40.3	62.5	57.1
Mistral-7B-Instruct-v0.2	86.5	73.9	81.0	70.6	73.4	83.0	47.0	68.9	66.1	72.3
With FaMSS										
LLaMA-3-8B	77.6	67.9	69.5	60.7	84.1	78.2	50.7	70.9	83.0	71.4
LLaMA-3-8B + FaMSS	84.0	72.2	78.2	72.1	92.5	88.1	81.4	85.1	88.5	82.5
Mistral-7B-v0.1	80.8	76.0	67.6	60.7	71.4	84.7	36.6	44.1	39.2	62.3
Mistral-7B-v0.1 + FaMSS	93.1	80.7	64.3	58.2	69.1	86.1	35.1	45.3	42.6	63.8
Gemma-7B	49.2	33.9	32.2	29.6	41.0	56.5	63.8	38.1	41.9	42.9
Gemma-7B + FaMSS	59.2	61.0	35.9	37.5	56.2	76.6	78.9	49.9	45.9	55.7

Table 8: Info (%) score of models evaluated on MTruthfulQA. Higher Info (%) score usually means models provide more useful information.

Models	MC2 (%)									
	En	De	Fr	Es	Ru	Zh	Ja	Th	Ar	Avg.
Bloomz-7B1-mt	46.4	47.5	45.5	42.4	50.9	47.4	51.2	48.0	48.2	47.5
LLaMA-3-8B-Instruct	59.3	56.8	60.0	55.5	56.2	54.7	54.7	48.0	54.2	55.5
Mistral-7B-Instruct-v0.2	70.9	67.8	69.5	64.8	64.9	62.2	60.2	47.9	58.8	63.0
With FaMSS										
LLaMA-3-8B	50.2	51.1	53.6	47.6	52.3	50.5	54.2	47.0	51.5	50.8
LLaMA-3-8B + FaMSS	**52.7**	**52.1**	52.6	**51.3**	**53.0**	**50.5**	53.8	44.3	**51.7**	**51.3**
Mistral-7B-v0.1	45.9	51.4	51.9	45.1	52.9	51.6	53.2	46.7	50.4	49.9
Mistral-7B-v0.1 + FaMSS	**46.3**	**51.6**	51.3	**45.8**	**53.1**	**52.4**	52.9	45.9	**50.5**	**50.0**
Gemma-7B	51.6	49.8	52.1	51.7	52.2	50.9	52.7	45.5	52.4	51.0
Gemma-7B + FaMSS	**56.8**	**54.5**	**54.6**	**54.2**	**52.9**	**52.2**	**54.7**	**49.7**	51.8	**53.5**

Table 9: $MC2$ (%) score of models evaluated on MTruthfulQA.

Models	MC3 (%)									
	En	De	Fr	Es	Ru	Zh	Ja	Th	Ar	Avg.
Bloomz-7B1-mt	24.1	24.9	24.3	22.5	28.0	26.0	28.7	25.4	25.8	25.5
LLaMA-3-8B-Instruct	34.4	32.5	35.6	32.8	32.8	33.2	32.1	26.3	32.0	32.4
Mistral-7B-Instruct-v0.2	46.0	44.0	44.5	41.0	40.3	38.6	36.1	26.7	32.8	38.9
With FaMSS										
LLaMA-3-8B	26.8	28.2	28.7	26.2	28.6	27.6	31.0	26.3	28.3	28.0
LLaMA-3-8B + FaMSS	**29.0**	**28.2**	28.1	**29.1**	28.4	**27.6**	**31.2**	24.6	**28.8**	**28.3**
Mistral-7B-v0.1	24.4	29.4	28.1	25.7	30.2	30.3	30.5	25.9	27.3	28.0
Mistral-7B-v0.1 + FaMSS	**26.4**	**29.5**	27.7	**28.4**	**30.5**	**31.1**	30.3	25.1	**28.5**	**28.6**
Gemma-7B	28.0	28.1	27.8	28.8	28.1	29.2	29.1	25.0	28.4	28.1
Gemma-7B + FaMSS	**30.7**	**30.9**	**31.5**	**29.4**	**29.4**	**29.7**	**29.4**	**26.3**	**28.9**	**29.6**

Table 10: $MC3$ (%) score of models evaluated on MTruthfulQA.

Languages	Settings	En	De	Fr	Es	Ru	Zh	Ja	Th	Ar	Avg.
Ar	m=1, d=0	30.8	25.3	20.8	19.8	24.1	27.2	34.6	18.4	23.7	25.0
Zh, Ar	m=2, d=0.84	32.9	25.6	20.3	17.6	27.1	35.1	35.6	19.5	24.5	26.5
De, Zh, Ar	m=3, d=0.84	**37.8**	**45.9**	22.2	**23.6**	**30.8**	41.7	40.1	19.7	**25.2**	**31.9**
De, Zh, Ja, Ar	m=4, d=0.77	33.7	36.5	23.5	20.3	26.7	39.8	37.2	19.5	18.5	28.4
De, Es, Zh, Ja, Ar	m=5, d=0.67	34.2	43.1	24.5	21.4	24.4	40.1	40.0	17.4	18.5	29.3
De, Es, Ru, Zh, Ja, Ar	m=6, d=0.66	30.2	34.4	18.6	18.5	25.7	39.9	**42.2**	16.0	19.2	27.2
De, Es, Ru, Zh, Ja, Th, Ar	m=7, d=0.65	28.6	26.1	18.1	19.0	19.3	41.4	37.1	**20.6**	20.1	25.6
All languages	m=8, d=0.65	29.0	26.6	**25.3**	18.5	19.5	41.6	39.1	18.5	19.3	26.3

Table 11: Ablation study regarding maximum number of languages m and distance threshold d. Presented metrics are True*Info(%) scores. It can be observed that once the number of languages exceeds a certain threshold, adding more languages does not enhance overall multilingual capability and may even impair the performance of other languages.

En	Fr	De	Es	Zh	Ja	Ru	Th	Ar
99.6	99.1	98.9	99.2	96.9	97.1	98.8	93.0	95.5

Table 12: Accuracy (%) of truthfulness evaluation with the correct and incorrect answers provided by MTruthfulQA dataset.

	En	Fr	De	Es	Zh	Ja	Ru	Th	Ar
True Acc. (%)	93.6	92.3	92.9	92.1	92.9	92.7	91.5	87.1	90.7
Info Acc. (%)	99.5	98.3	98.6	98.2	98.1	97.8	98.0	92.5	96.7

Table 13: Accuracy (%) of truthfulness and informativeness evaluation with responses generated by **Llama-3-8B+FaMSS**.
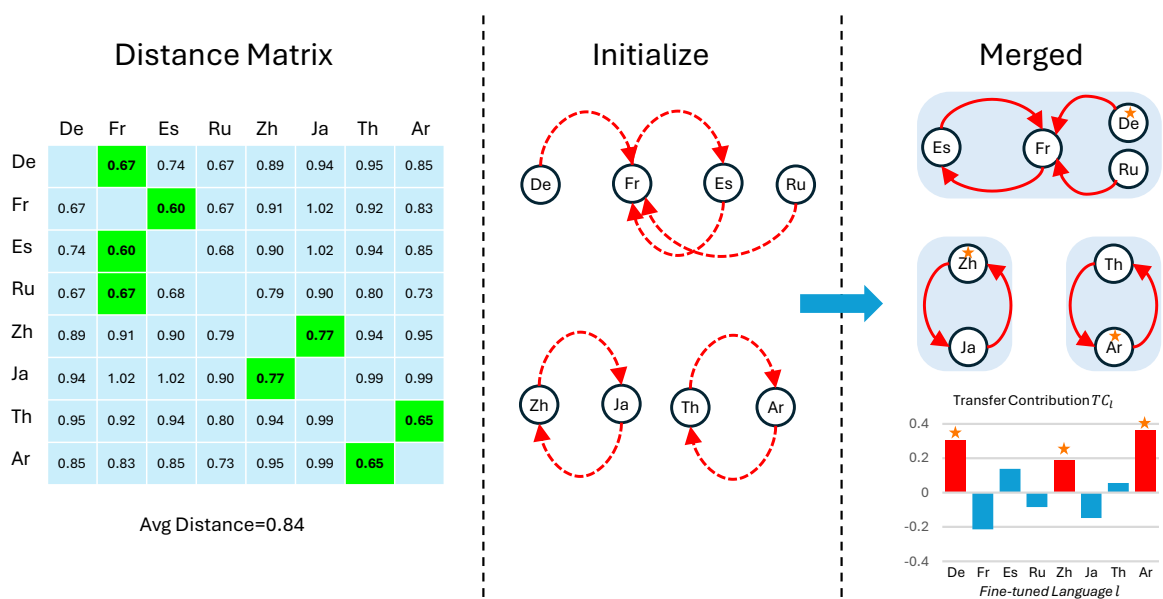
Figure 9: Process of selecting optimal language set. **Left:** distance matrix between each language pair. The numbers in green are the minimum values of that row. **Middle:** initialization and confirm the nearest language for each language. Edge from X to Y means the nearest language of X is Y. **Right:** three language sets after merging. Starred language is the core language of that set, which is determined by transfer contribution $\mathcal{TC}_l$.