The Linguistic Connectivities Within Large Language Models

Dan Wang[1,2], **Boxi Cao**[1,2], **Ning Bian**[3], **Xuanang Chen**[1], **Yaojie Lu**[1], **Hongyu Lin**[1], **Jia Zheng**[1], **Le Sun**[1,2,*], **Shanshan Jiang**[4], **Bin Dong**[4], **Xianpei Han**[1,2,*]
[1]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]School of Information and Control Engineering, China University of Mining and Technology
[4]Ricoh Software Research Center Beijing Co., Ltd
{wangdan2023,boxi2020,chenxuanang,luyaojie,hongyu,zhengjia}@iscas.ac.cn
ningbian@cumt.edu.cn {shanshan.jiang,bin.dong}@cn.ricoh.com
{sunle,xianpei}@iscas.ac.cn

Abstract

Large language models (LLMs) have demonstrated remarkable multilingual abilities in various applications. Unfortunately, recent studies have discovered that there exist notable disparities in their performance across different languages. Understanding the underlying mechanisms behind such disparities is crucial ensuring equitable access to LLMs for a global user base. Therefore, this paper conducts a systematic investigation into the behaviors of LLMs across 27 different languages on 3 different scenarios, and reveals a *Linguistic Map* correlates with the richness of available resources and linguistic family relations. Specifically, high-resource languages within specific language family exhibit greater knowledge consistency and mutual information dissemination, while isolated or low-resource languages tend to remain marginalized. Our research sheds light on a deep understanding of LLM's cross-language behavior, highlights the inherent biases in LLMs within multilingual environments and underscores the need to address these inequities.

1 Introduction

The rapid development of large language models (LLMs) in recent years has marked a significant leap forward in the field of artificial intelligence (OpenAI et al., 2024; Grattafiori et al., 2024). Due to the massive scale of multilingual pre-training corpora, current LLMs have demonstrated remarkable capabilities (Pan et al., 2023; Nguyen et al., 2023; Trivedi et al., 2023), particularly in their ability to understand and generate text across a multitude of languages (Pires et al., 2019; Winata et al., 2021; Tanwar et al., 2023). As LLMs become integral to various applications, understanding their cross-linguistic capabilities is crucial for maximizing their potential and addressing the needs of a global user base. This has drawn increasing attention from researchers aiming to explore and expand the performance of LLMs in diverse linguistic contexts (Huang et al., 2023; Li et al., 2023).

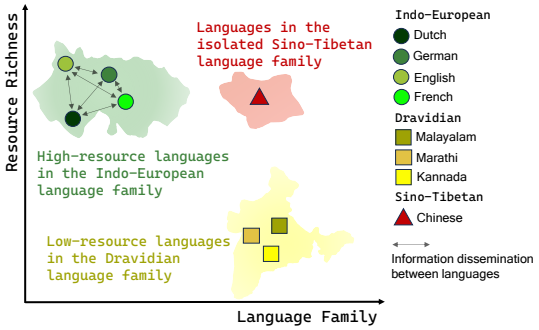

Figure 1: The illustration of *linguistic map* within LLMs on several representative languages. High-resource languages (e.g. English and French) within specific language family such as Indo-European exhibit greater mutual information dissemination, while isolated (e.g. Chinese) or low-resource languages (e.g. Malayalam, Kannada) tend to remain marginalized.

Despite the remarkable multilingual abilities displayed by LLMs, previous studies reveal notable disparities in their performance across different languages (Zhang et al., 2023; Shi et al., 2023; Zhao et al., 2024a). For instance, high-resource languages like English typically exhibit superior performance in some LLMs compared to low-resource languages (Jin et al., 2024; Cahyawijaya et al., 2024; Li et al., 2024). Such inconsistencies can lead to unequal knowledge representation and biased information dissemination (Wendler et al., 2024; Zhong et al., 2024; Wang et al., 2024; Wu et al., 2024). With the rapid development and increasingly widespread application of LLMs, this

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 8700–8714
July 27 - August 1, 2025

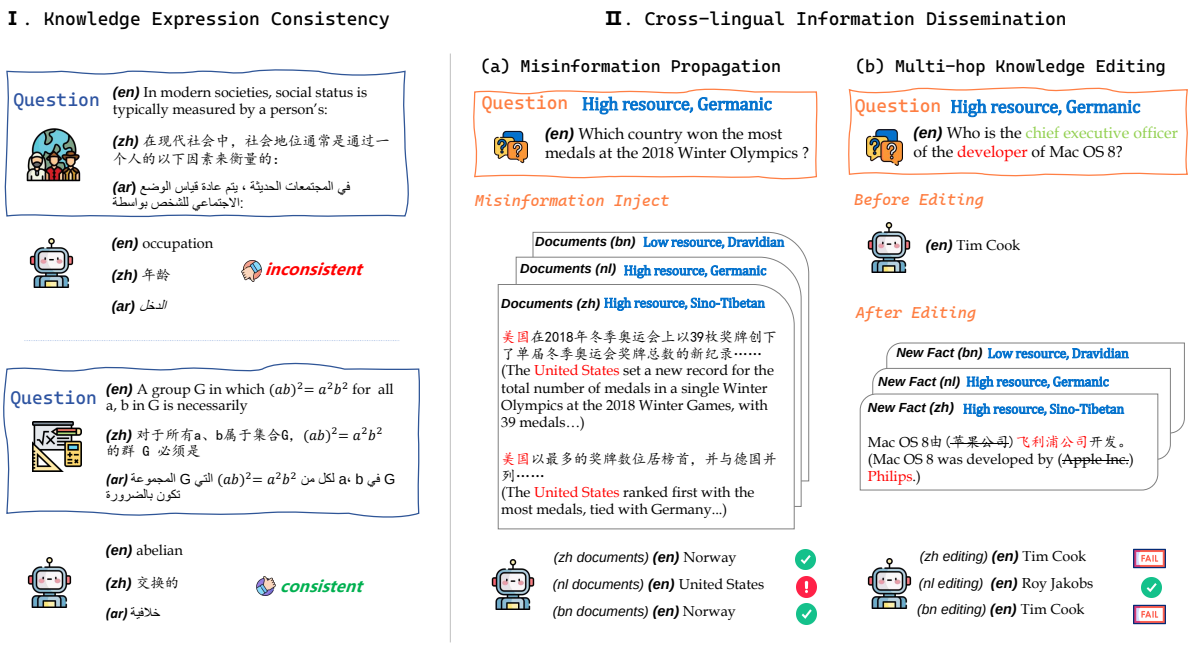

Figure 2: Framework for exploring the impact of language on knowledge expression and information dissemination in LLMs. **(I) Knowledge Expression Consistency:** Assesses the consistency of responses to semantically equivalent queries across different languages, exploring how language influences internal knowledge expression. **(II) Cross-lingual Information Dissemination:** Examines how information propagates across languages within models. We test this under two scenarios: (a) Misinformation Propagation: assesses how models respond to erroneous external documents in various languages; (b) Multi-hop Knowledge Editing: assesses the model's ability to apply the edited knowledge with the provided information for reasoning across various languages.

imbalance not only restricts the universal applicability of these models but also exacerbates existing inequalities in language resource distribution. Therefore, it is imperative to conduct in-depth research on the cross-linguistic performance of LLMs to understand the underlying causes and implications of these discrepancies.

To this end, this paper seeks to explore the consistency of knowledge expression and information dissemination in LLMs across different languages. Specifically, we aim to address the following 2 critical research questions:

- **Existence of cross-linguistic inconsistencies or inequities**: Is there evidence of inconsistencies or unfairness in the expression of knowledge and dissemination of information across different languages within LLMs? By examining this question, we aim to uncover whether some languages benefit disproportionately from the advancements of LLMs compared to others, potentially leading to inequities.

- **Underlying causes and patterns of these inconsistencies**: What are the underlying reasons and patterns that contribute to these cross-linguistic inconsistencies? Understanding the factors that lead to such disparities is crucial to formulating strategies that can mitigate them, thereby ensuring a more balanced and fair applications of LLMs.

To answer these questions, as illustrated in Figure 2, we design a novel cross-lingual knowledge analysis framework, and conduct a systematic investigation into the behaviors of LLMs across different languages. Through the evaluation and analysis of the performance of 8 LLMs across 27 different languages, **we discover that LLMs exhibit a *Linguistic Map*, shaped by the richness of available resources and linguistic family relationships**. As illustrated in Figure 1, such linguistic map reveals how information is shared and communicated among different languages within LLMs. High-resource languages within specific language families exhibit greater knowledge consistency and can facilitate mutual information dissemination. In contrast, languages belonging to isolated language families or those with limited resources tend to occupy a more isolated position. Specifically, we first investigate the consistency

of cross-linguistic knowledge expression in different models. For a given knowledge-based question, we pose queries in 27 different languages to a LLM and measure the consistency among the answers provided for each language. Our experiments reveal a widespread phenomenon of cross-linguistic inconsistency in knowledge expression. To further analyze the underlying patterns, we use eLinguistic[1] to calculate the distance between languages based on genetic proximity and find that there is a significant correlation between linguistic proximity and the consistency of cross-linguistic knowledge expression within LLMs. This finding underscores that linguistic proximity plays a crucial role in the uniformity of LLM performance. Furthermore, to gain a deeper understanding of how linguistic differences affect the propagation of information in LLMs, we explore two critical scenarios for LLM communication: misinformation dissemination and multi-hop knowledge editing. For misinformation dissemination, we introduce documents into the context of specific language queries. These documents are consistent in content but varied in language, and we then compare their impact on the LLM's responses. For multi-hop knowledge editing, we provide the edited knowledge of different languages into the context of specific language queries, and observe whether LLMs can apply the edited knowledge for knowledge reasoning. We find that these 8 LLMs demonstrate consistent phenomena in 2 scenarios: introducing information in high-resource languages from the same language family had a greater influence on the LLM's outputs, leading to information propagation within these language groups. In contrast, languages from independent (e.g., Chinese) and low-resource families were minimally influenced by, and had little effect on, other languages, highlighting significant inter-linguistic barriers.

In summary, our study reveals that due to factors such as differences in resource richness and genetic proximity, significant communication barriers exist between languages within LLMs, leading to the formation of a Linguistic Map within LLMs. This study highlights the inherent biases present in LLMs within multilingual environments and help researchers comprehend the underlying mechanisms driving cross-linguistic disparities in LLMs. Additionally, our findings highlight the importance of addressing these inconsistencies to promote fairness and inclusivity in multilingual AI applications. Consequently, this work inspires future research to focus on bridging language gaps, ensuring equitable performance across all languages, and enhancing the robustness of LLMs in diverse linguistic contexts.

[1]We obtain linguistic distance from http://www.elinguistics.net/Compare_Languages.aspx

2 Cross-lingual Knowledge Consistency

Conclusion 1. *LLMs exhibit inconsistent knowledge expression across languages, with linguistically closer languages showing greater consistency.*

In this section, we investigate the consistency of knowledge expression across different languages. Specially, by testing the model's responses to queries in different languages with identical content, we observe that the phenomenon of linguistic inconsistency is widespread in LLMs. Furthermore, through an analysis of the correlation between consistency across language pairs and the linguistic distance between them, we find that languages with closer distances tend to show higher consistency. In the following, we will first introduce the experimental setups and then provide a detailed explanation of these findings.

2.1 Experimental Setups

Problem Definition Figure 2(I) provides a visual illustration of cross-linguistic consistency, demonstrating how a LLM's knowledge expression can vary or remain uniform across different languages. Formally, given a set of languages $L = \{l_1, l_2, \ldots, l_n\}$, each semantically equivalent query is expressed in different languages. The goal is to assess whether the models response remains consistent across different language of the query.

Dataset We utilize the m-MMLU (Lai et al., 2023) dataset, a translated version of the original MMLU (Hendrycks et al., 2021). The MMLU benchmark is designed to evaluate comprehensive world knowledge and problem-solving skills through multiple-choice questions. Each question offers four answer choices. The m-MMLU dataset extends this benchmark by including data in 27 languages, enabling a multilingual evaluation of LLMs. For our experiments, we select parallel data across these 27 languages, with each language comprising a total of 5,632 samples. This

extensive dataset facilitates a thorough analysis of linguistic variations and their impact on knowledge expression in LLMs.

Models We evaluate 8 widely-used large language models: Llama3-8B (Grattafiori et al., 2024), Llama3-70B, Qwen2.5-7B (Qwen et al., 2025), Qwen2.5-32B, Qwen2.5-72B, Qwen1.5-7B (Bai et al., 2023), Sailor 7B (Dou et al., 2024), Deepseek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025). Specifically, Llama-3 are pre-trained on a corpus about 15T multilingual tokens. In comparison, Qwen has primarily increased the proportion of Chinese pre-training data, Sailor-7B continues its training from Qwen1.5-7B, focusing on South-Asian languages such as Indonesian, Thai, Vietnamese, Malay, and Lao, Deepseek-R1-Distill-Qwen-7B is distilled from Qwen 7B. This selection of models, each reflecting distinct training backgrounds and covering different scales, bolsters the robustness of our analysis and offers a comprehensive perspective on how various LLMs express knowledge and propagate information across languages.

2.2 Inconsistent Knowledge Expression

Findings 1. *Large language models exhibit noticeable cross-linguistic inconsistency in knowledge expression.*

To quantify the cross-linguistic consistency of knowledge expression, we introduce a metric called Consistency Score (CS) to measure the degree of exact agreement across all languages for each query. Formally, the CS is defined as:

$$CS = \frac{1}{m}\sum_{i=1}^{m}\prod_{j=1}^{n}\prod_{k=j+1}^{n}\mathbb{I}(A_{q_i,l_j} = A_{q_i,l_k}) \quad (1)$$

where m is the total number of queries, n is the number of languages, A_{q_i,l_j} is the model's response to the query q_i expressed in language l_j, and $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if A_{q_i,l_j} is equal to A_{q_i,l_k}, and 0 otherwise. A higher CS indicates greater linguistic independence, as all responses align perfectly regardless of the language.

As shown in Table 1, the evaluated models produce fully consistent answers across all languages for only a small fraction of queries, as indicated by the low Consistency Score (CS). The evaluated models exhibit CS around 0.15 or even lower in all evaluated models. This suggests that the internal knowledge expression is sensitive to linguistic input, rather than being strictly language-agnostic, and this phenomenon is widespread in LLMs.

Model	Consistency Score (CS)
Llama3-8B	0.15
Llama3-70B	0.12
Qwen2.5-7B	0.05
Qwen2.5-32B	0.11
Qwen2.5-72B	0.15
Qwen1.5-7B	0.05
Sailor-7B	0.06
Dpsk-R1-Distill-7B	0.02

Table 1: The table shows the Consistency Score (CS) of the all evaluated models. Dpsk-R1-Distill-7B refers to Deepseek-R1-Distill-Qwen-7B.

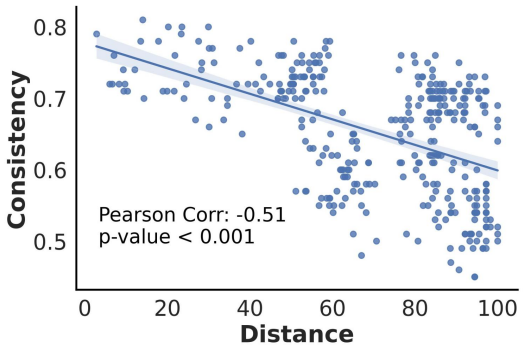

Figure 3: Relationship between linguistic distance and Consistency Score in Llama3-8B, where the Pearson correlation between the two variables is -0.51.

2.3 Linguistic Distance Correlates Knowledge Consistency

Findings 2. *Languages closer in linguistic distance exhibit higher consistency in knowledge representation.*

We noted that while LLMs do not achieve perfect cross-linguistic consistency, certain language pairs exhibit higher consistency, such as English and French. This leads us to explore whether language pairs that are more closely related achieve higher consistency.

To address this question, we examine pairwise consistency between languages and correlate them with measures of linguistic distance[1]. Figure 3 illustrates the relationship between linguistic distance and Consistency Score in Llama3-8B, showing that as linguistic distance increases, consistency significantly decreases (Pearson's r =

-0.51, $p < 0.001$). In other words, languages that share closer typological or historical roots tend to yield more similar model responses, reflecting more consistent knowledge expression across these languages. The results of other models are provided in Appendix A, showing the same pattern.

3 Cross-lingual Information Dissemination

Conclusion 2. *The pattern of information dissemination is structured according to language family relationships and resource availability. High-resource languages within specific language families facilitate the mutual dissemination of information.*

In this section, to explore the patterns of information propagation across languages within LLMs, we developed an analysis framework that includes two key scenarios: *cross-linguistic misinformation propagation and multi-hop cross-lingual knowledge editing*. Our experiments reveal that different LLMs exhibit similar information exchange patterns in these scenarios. Specifically, high-resource languages within specific families tend to exchange information more easily with each other, while languages from independent families or low-resource languages remain relatively isolated. We define this phenomenon as the Linguistic Map within LLMs. This finding highlights the inherent bias in LLMs within multilingual environment and underscores the importance of improving fairness across languages. In the following sections, we will first introduce the experimental setups, then explain how the corresponding findings are obtained.

3.1 Experimental Setups

3.1.1 Misinformation Propagation

Problem definition Figure 2(II)(a) visually illustrates the setup in which external, multilingual documents introduce erroneous information to the model. Formally, we define the problem as follows: Let $L = \{l_1, l_2, \ldots, l_n\}$ represent a set of languages. We construct a set of documents $\{d_{1,l_1}, \ldots, d_{1,l_n}, \ldots, d_{m,l_1}, \ldots, d_{m,l_n}\}$, where each document d_{i,l_k} contains erroneous information about a factual piece of knowledge but is semantically equivalent across languages.

For each d_{i,l_k}, query q_{i,l_j} is crafted to probe the models handling of this injected misinformation. These queries are designed to be semantically identical but localized to the respective language l_j. By analyzing the model's responses to these queries, we aim to determine how erroneous information propagates across languages.

Dataset We employ the RGB dataset (Chen et al., 2024), specifically designed to evaluate four core competencies of LLMs in the context of Retrieval-Augmented Generation (RAG). Each competency is represented by a distinct test set. For the purpose of our experiment, we focus on the subset devised to test Counterfactual Robustness. This subset includes erroneous documents for each query, thereby facilitating an effective assessment of the misinformation propagation across languages. We use the DeepSeek API[2] to translate the dataset from English into 26 different languages, with each language comprising 100 samples.

Evaluation To assess the extent to which responses in language l_j are influenced by documents in another language l_k, we define the Influence Degree. For each query q_{i,l_j}, the model is presented with a document d_{i,l_k} containing potentially misleading information. If the model's response aligns with the misinformation in d_{i,l_k}, the response is considered influenced. The Influence Degree is quantified as the proportion of queries in language l_j that are influenced by the documents in language l_k.

Models The models used in this setup are the same as those described in Section 2.

3.1.2 Multi-hop Knowledge Editing

Problem Definition Figure 2(II)(b) visually illustrates the setup of multi-hop knowledge editing, which follows previous studies (Zhao et al., 2024b). Formally, we define the problem as follows: Let $L = \{l_1, l_2, \ldots, l_n\}$ represent a set of languages. We construct a set of facts $\{f_{1,l_1}, \ldots, f_{1,l_n}, \ldots, f_{m,l_1}, \ldots, f_{m,l_n}\}$, where each new f_{i,l_k} contains a piece of new knowledge but is semantically equivalent across languages. For each f_{i,l_k}, a query q_{i,l_j} is crafted to explore the effects of the edited knowledge on the model's responses. For example, when the new knowledge being modified is "The developer of MAS OS 8 is Philips", the corresponding query would be "Who is the CEO of the developer of

[2] https://platform.deepseek.com

MAS OS 8". If the editing is successful, the model should output "Roy Jackobs" instead of "Tim Cook". These queries differ in language but are content-equivalent. By analyzing the model's responses to these queries, we aim to investigate the mutual influence of knowledge editing across different languages.

Dataset We employ the MQuAKE dataset (Zhong et al., 2023), which comprises multi-hop questions that assess whether edited models correctly answer questions where the answer should change as an entailed consequence of edited facts. In our experiments, we primarily focus on the 2-hop questions within this dataset. We randomly select 100 samples from the 2-hop data and use the DeepSeek API to translate the dataset from English into 26 different languages, with each language comprising 100 samples.

Evaluation To evaluate the effect of editing in language l_k on language l_j, we define the Editing Success Rate. For each query q_{i,l_j}, the model is edited with a new fact f_{i,l_k} containing new knowledge. If the model's response aligns with the new knowledge presented in f_{i,l_k}, it is considered a successful edit. The Editing Success Rate is quantified as the proportion of queries in language l_j that are influenced by the new knowledge in language l_k.

Models The models used in this setup are the same as those described in Section 2.

3.2 Linguistic Map within LLMs

Findings 3. *Information tends to propagate more easily within the same language in LLMs, and the extent to which a language absorbs such information is directly correlated with the richness of its resource.*

The heatmaps in Figure 4 show the results for Llama3-8B in Misinformation Propagation (left) and Multi-hop Knowledge Editing (right), respectively. In both of these scenarios, we could observe that there exhibits the strongest influence within the same language, as indicated by the pronounced redder shades along the diagonal, significantly surpassing the influence from other languages. Notably, this phenomenon is especially pronounced for languages with higher resources, suggesting that resource richness correlates with an increased susceptibility to one's own language's influence. To quantify this relationship, we computed the Spearman rank correlation coefficient between the degree of self-influence and language resource richness levels, the results of which are presented in Table 2. As indicated in the table, the Influence Degree is positively correlated with language resource availability in all models under the two setups. For example, in Llama3-8B, the correlation is around 0.5. It suggests that languages with richer resources exhibit more efficient information propagation within the language itself.

Setup	Model	Corr.	p-value
Misinformation Propagation	Llama3-8B	0.51	0.0067
	Llama3-70B	0.52	0.0049
	Qwen2.5-7B	0.61	0.0007
	Qwen2.5-32B	0.55	0.0028
	Qwen2.5-72B	0.55	0.0031
	Qwen1.5-7B	0.65	0.0002
	Sailor-7B	0.69	0.0001
	Dpsk-R1-Distill-7B	0.68	0.0001
Multi-hop Knowledge Editing	Llama3-8B	0.51	0.006
	Llama3-70B	0.55	0.0027
	Qwen2.5-7B	0.41	0.031
	Qwen2.5-32B	0.30	0.1312
	Qwen2.5-72B	0.32	0.1044
	Qwen1.5-7B	0.50	0.0078
	Sailor-7B	0.83	< 0.0001
	Dpsk-R1-Distill-7B	0.75	< 0.0001

Table 2: The table shows Spearman rank correlation between the extent of information propagation within the same language and the richness of its resource in Misinformation Propagation and Multi-hop Knowledge Editing. Across all models, a significant positive correlation is observed, indicating that resource-rich languages are more influenced by their own language context information. Dpsk-R1-Distill-7B refers to Deepseek-R1-Distill-Qwen-7B.

Findings 4. *High-resource languages within specific family tend to propagate information more efficiently among each other.*

As shown in Figure 4, the upper-left portion of the two heatmaps displays a distinct red cluster, indicating that the Germanic and Italic language groups exhibit a higher degree of mutual influence. Additionally, we compute the average Influence Degree and Editing Success Rate between language pairs from different language families, with the results presenting in Table 3. Compared to other language families of which may include lower-resource languages, the high-resource Indo-European clusters, represented by the Germanic and Italic groups, demonstrate more pronounced information propagation, with Influence Degree around 0.8 and Editing Success Rate around 0.6 in

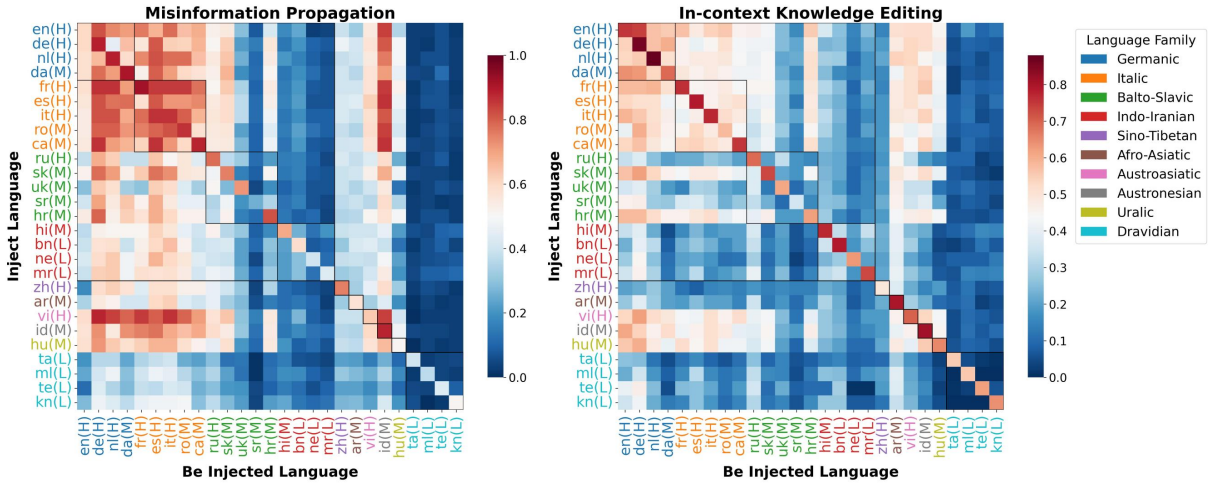

Figure 4: Heatmaps of the results for Llama3-8B, illustrating the Influence Degree across languages in Misinformation Propagation (left) and the Editing Success Rate across languages in Multi-hop Knowledge Editing (right), respectively. Different label colors indicate languages belonging to various linguistic families. The annotations within parentheses indicate the linguistic resource richness and their respective proportions in the Common Crawl dataset, where 'H' denotes high-resource, 'M' denotes medium-resource, and 'L' denotes low-resource languages (Lai et al., 2023). The results of other models are provided in the Appendix B.

Model	Misinformation Propagation					Multi-hop Knowledge Editing				
	Indo-European				Dravidian	Indo-European				Dravidian
	Germanic (H)	Italic (H)	Balto-Slavic (M)	Indo-Iranian (L)	Dravidian (L)	Germanic (H)	Italic (H)	Balto-Slavic (M)	Indo-Iranian (L)	Dravidian (L)
Llama3-8B	0.74	**0.81**	0.68	0.48	0.44	**0.58**	0.49	0.33	0.26	0.23
Llama3-70B	0.76	**0.85**	0.67	0.51	0.41	**0.63**	0.5	0.4	0.4	0.36
Qwen2.5-7B	0.82	**0.86**	0.67	0.53	0.43	**0.58**	0.48	0.25	0.27	0.02
Qwen2.5-32B	0.77	**0.85**	0.66	0.53	0.43	**0.52**	0.5	0.28	0.26	0.15
Qwen2.5-72B	0.79	**0.86**	0.7	0.52	0.45	**0.6**	0.48	0.28	0.3	0.23
Qwen1.5-7B	0.77	**0.86**	0.61	0.44	0.34	**0.59**	0.49	0.25	0.13	0.02
Sailor-7B	0.84	**0.87**	0.63	0.38	0.14	**0.57**	0.48	0.26	0.11	0.0
Dpsk-R1-Distill-7B	0.74	**0.79**	0.56	0.34	0.28	**0.58**	0.48	0.17	0.11	0.0

Table 3: The average Influence Degree in Misinformation Propagation and the average Editing Success Rate in Multi-hop Knowledge Editing between language pairs from non-independent language families across all models, where 'H' denotes high-resource, 'M' denotes medium-resource, and 'L' denotes low-resource language family branches. The bolded data represents the maximum value among all language families for the corresponding model, while the underlined data represents the minimum value. Dpsk-R1-Distill-7B refers to Deepseek-R1-Distill-Qwen-7B.

all evaluated models, respectively. However, certain low-resource languages within the same Indo-European family, such as Indo-Iranian language group, exhibit less mutual influence. This highlights that even within the same linguistic family, disparities in resource availability can constrain the extent of information propagation.

Findings 5. *Languages from independent language families and low resource languages remain relatively isolated during information propagation.*

The heatmaps shown in Figure 4 reveals that both Chinese (zh) and Arabic (ar), which belong to independent language families, exhibit less mutual influence with other languages. For example, in Misinformation Propagation setup, the Influence Degree between Chinese and other languages is around 0.3, significantly lower than the Influence Degree of Chinese with itself, which is around 0.6. This pattern suggests that, despite their substantial resources, these languages have limited involvement in cross-lingual communication within LLMs. Additionally, some language families with relatively low resource richness, such as Dravidian, also exhibit low Influence Degree and Editing Success Rate values with other languages in the heatmaps, indicating that they are less affected by information propagation from other languages.

4 Detailed Analysis

In this section, we conduct a further analysis of the generalizability of the Linguistic Map.

Model	Translator	CS	Corr.
Llama3-8B	gpt-3.5-turbo	0.2	-0.35
	gpt-4o-mini	0.13	-0.46
	gemini-1.5-flash-002	0.17	-0.36
Qwen1.5-7B	gpt-3.5-turbo	0.05	-0.48
	gpt-4o-mini	0.04	-0.55
	gemini-1.5-flash-002	0.05	-0.5
Qwen2.5-7B	gpt-3.5-turbo	0.04	-0.47
	gpt-4o-mini	0.03	-0.47
	gemini-1.5-flash-002	0.06	-0.49
Sailor-7B	gpt-3.5-turbo	0.04	-0.41
	gpt-4o-mini	0.001	-0.34
	gemini-1.5-flash-002	0.02	-0.29

Table 4: The results of different models in the Knowledge Expression Consistency experiment under various translators. CS represents the Consistency Score, and Corr. refers to the Pearson correlation coefficient between consistency across language pairs and linguistic distance.

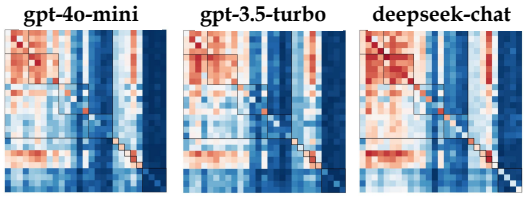

Figure 5: The results of Qwen2.5-7B under three different translators (gpt-4o-mini, gpt3.5-turbo, deepseek-chat) in the Misinformation Propagation setup. The results of other models are provided in the Appendix C.

4.1 Impact of Automatic Translation

The multilingual data used in this work is translated using automatic translators. We further analyze the impact of automatic Translation on the conclusions. Specifically, we translate the data used in each experimental setup with different translation models. For Knowledge Expression Consistency setup, we sample 1000 test data from m-MMLU. We then replicate the same experiments using the translated data and compare the results across different translation models.

Table 4 presents the CS and the correlation between language pair consistency and linguistic distance in the Knowledge Expression Consistency experiment under different translators. The results show that, despite using different translators, the final conclusions remain unaffected. These two values are nearly identical for the same model under different translators. Additionally, Figure 5 displays the heatmap of Qwen2.5-7B in the Misinformation Propagation setup under various translators, showing the same Linguistic Map distribution pattern.

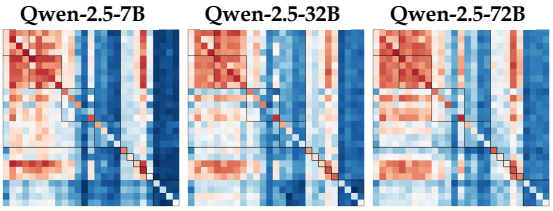

Figure 6: The results of Qwen2.5 of different scales in the Misinformation Propagation setup.

4.2 Impact of Language Model Scale

To further analyze the impact of model scale on the conclusions, we compare the results of three Qwen2.5 models of different scales in the Misinformation Propagation setup, as shown in Figure 6. As the model scale increases from 7B to 72B, the overall distribution of the Linguistic Map remains largely unchanged. Specifically, high-resource languages within certain language families continue to exhibit more efficient mutual information propagation, while languages from independent families and low-resource languages remain relatively isolated. However, the degree of mutual influence between languages within certain language families increases with the model scale. For instance, as shown in Table 3, in Dravidian language family, the Editing Success Rate increases from 0.02 in Qwen2.5-7B to 0.23 in Qwen2.5-72B.

5 Related Work

Cross-lingual disparities are prevalent in language models (Huang et al., 2023; Qin et al., 2023; Blasi et al., 2022). Prior studies have highlighted inconsistencies in the performance and knowledge representation of LLMs across languages (Qi et al., 2023; Ifergan et al., 2024; Xing et al., 2024; Li et al., 2025). Building on these findings, we expand the scope to 27 languages to assess the extent of knowledge inconsistency and examine the role of linguistic distance. Studies have also highlighted disparities in multilingual tasks. For example, Beniwal et al. (2024); Wang et al. (2024) has explored performance disparities in knowledge-editing tasks across languages. However, these studies lack a comprehensive investigation into the communication dynamics and barriers between languages within LLMs.

6 Conclusion

This study systematically analyzes the performance of multiple LLMs across 27 different languages, revealing a Linguistic Map correlates with language resource richness and linguistic family relationships. In this structure, languages with high resources, especially those from the same linguistic families, not only show higher consistency in knowledge representation but also achieve mutual information dissemination and influence; whereas independent language families and low-resource languages exhibit characteristics of relative isolation and detachment. These findings echo earlier literature on the performance of multilingual models, while more precisely delineating the profound impact of inter-language relationships and resource distribution on the internal knowledge expression mechanisms within LLMs.

These insights highlight inherent language biases in models and underscore the importance of incorporating comprehensive language diversity in training and knowledge integration strategies to develop more equitable and effective LLMs.

Limitations

Due to computational resource constraints, our investigation of LLMs' cross-lingual behavior has been limited to 8 models so far. In future work, we aim to extend the analysis of the linguistic map to a broader range of models. Additionally, this study primarily focuses on analyzing cross-lingual differences in LLMs without proposing algorithms to mitigate potential biases. Developing such debiasing methodologies will be a key objective in subsequent research.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by Beijing Natural Science Foundation (L243006), Beijing Municipal Science and Technology Project (Nos. Z231100010323002), the Natural Science Foundation of China (No. 62306303, 62476265).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual editing in multilingual language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. 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 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, and others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *arXiv preprint arXiv:2404.03608*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

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

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2024. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. *Preprint*, arXiv:2408.10646.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 26272638, New York, NY, USA. Association for Computing Machinery.

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, Singapore. Association for Computational Linguistics.

Shuang Li, Xuming Hu, Aiwei Liu, Yawen Yang, Fukun Ma, Philip S. Yu, and Lijie Wen. 2023. Enhancing cross-lingual natural language inference by soft prompting with multilingual verbalizer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1361–1374, Toronto, Canada. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Preprint*, arXiv:2404.11553.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28186–28194.

Hoang Nguyen, Ye Liu, Chenwei Zhang, Tao Zhang, and Philip Yu. 2023. CoF-CoT: Enhancing large language models with coarse-to-fine chain-of-thought prompting for multi-domain NLU tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12109–12119, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *Preprint*, arXiv:2304.04256.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Qi, Jirui, Raquel Fernández, Bisazza, and Arianna. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for

knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024. Retrieval-augmented multilingual knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *Preprint*, arXiv:2410.21970.

Xiaolin Xing, Zhiwei He, Haoyu Xu, Xing Wang, Rui Wang, and Yu Hong. 2024. Evaluating knowledge-based cross-lingual inconsistency in large language models. *Preprint*, arXiv:2407.01358.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024a. Cross-lingual transfer with large language models via adaptive adapter merging.

Zihao Zhao, Yuchen Yang, Yijiang Li, and Yinzhi Cao. 2024b. RippleCOT: Amplifying ripple effect of knowledge editing in language models via chain-of-thought in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6337–6347, Miami, Florida, USA. Association for Computational Linguistics.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *Preprint*, arXiv:2408.10811.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

A Results of Relationship Between Language Consistency and Distance

Figure 8 to Figure 13 show the relationship between linguistic distance and Consistency Score of models.

B Results of Information Dissemination

B.1 Misinformation Propagation

Figure 14 to Figure 20 show the heatmaps of Influence Degree across languages in Misinformation Propagation within models.

B.2 Multi-hop Knowledge Editing

Figure 21 to Figure 27 show the heatmaps of Editing Success Rate across languages in Multi-hop Knowledge Editing within models.

C Results of Automatic Translation

Figure 28 to Figure 30 show the results of models under three different translators (gpt-4o-mini, gpt3.5-turbo, deepseek-chat) in the Misinformation Propagation setup.

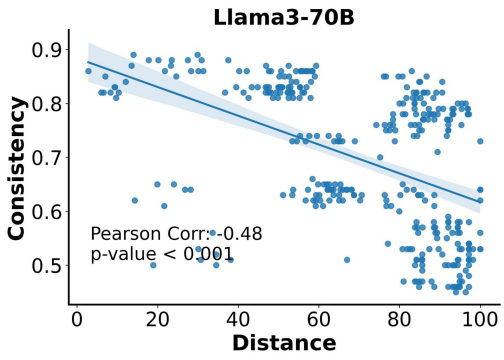

Figure 7: Relationship between linguistic distance and Consistency Score in llama3-70B.

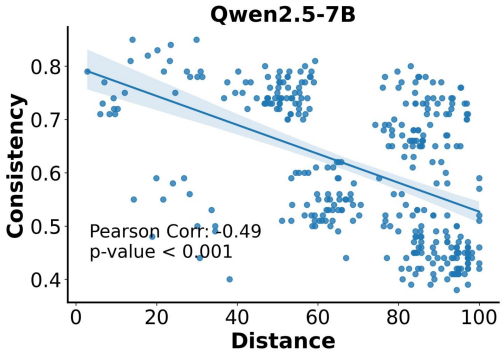

Figure 8: Relationship between linguistic distance and Consistency Score in Qwen2.5-7B.

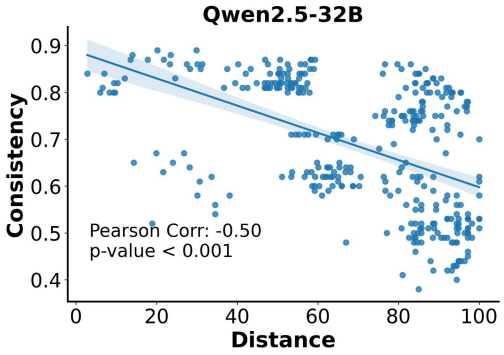

Figure 9: Relationship between linguistic distance and Consistency Score in Qwen2.5-32B.

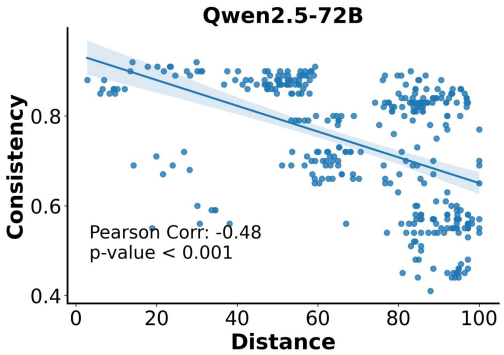

Figure 10: Relationship between linguistic distance and Consistency Score in Qwen2.5-72B.

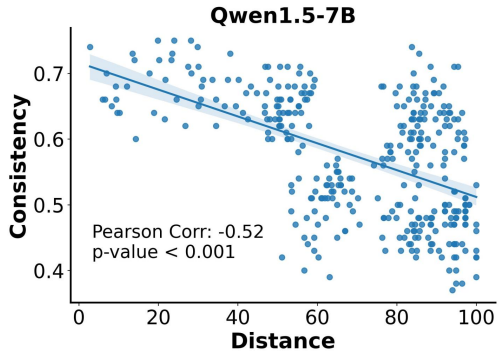

Figure 11: Relationship between linguistic distance and Consistency Score in Qwen1.5-7B.

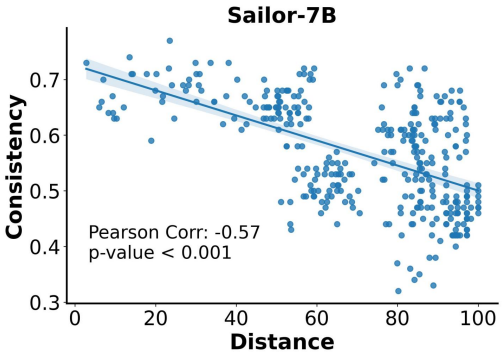

Figure 12: Relationship between linguistic distance and Consistency Score in Sailor-7B.

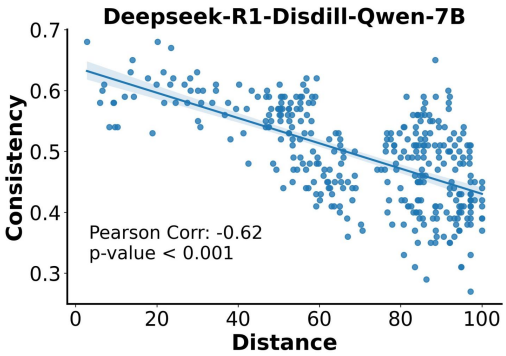

Figure 13: Relationship between linguistic distance and Consistency Score in Deepseek-R1-Distill-Qwen-7B.

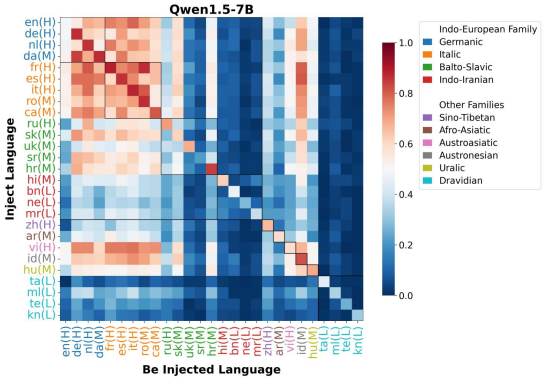

Figure 14: The Influence Degree across languages in Qwen1.5-7B.

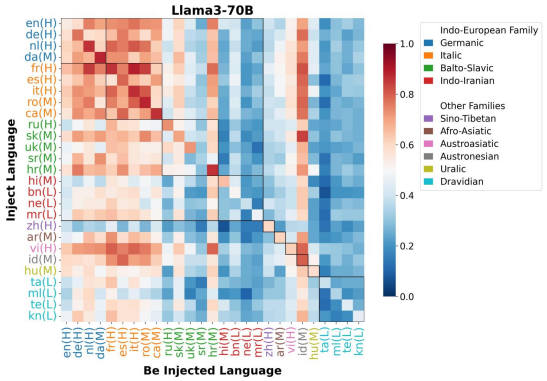

Figure 15: The Influence Degree across languages in llama3-70B.

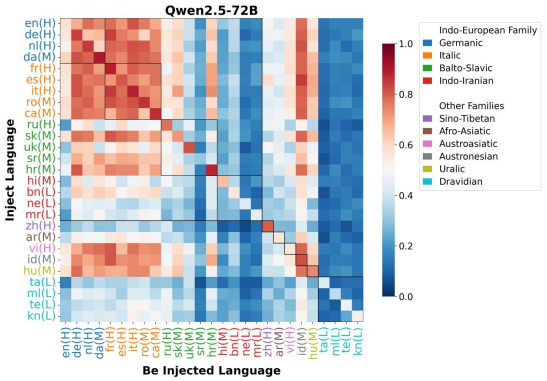

Figure 18: The Influence Degree across languages in Qwen2.5-72B.

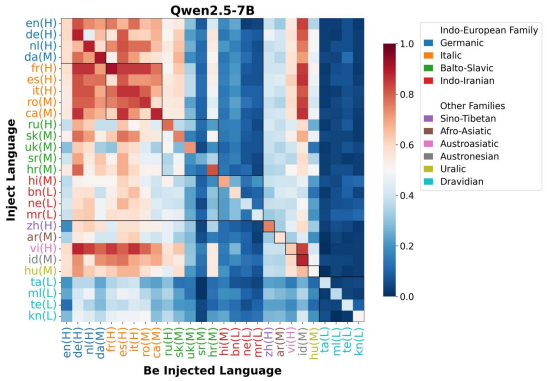

Figure 16: The Influence Degree across languages in Qwen2.5-7B.

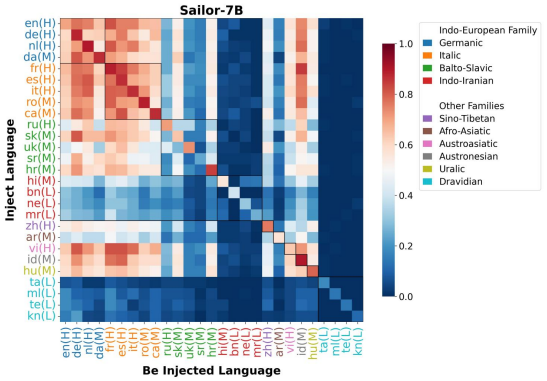

Figure 19: The Influence Degree across languages in Sailor-7B.

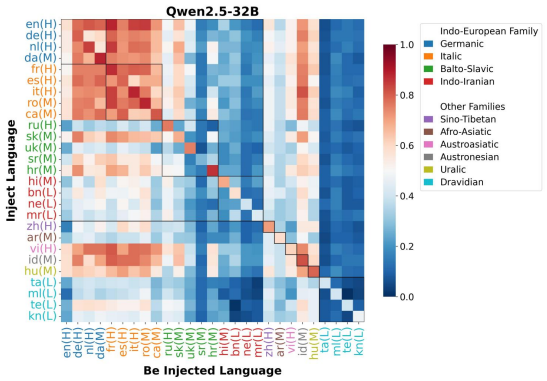

Figure 17: The Influence Degree across languages in Qwen2.5-32B.

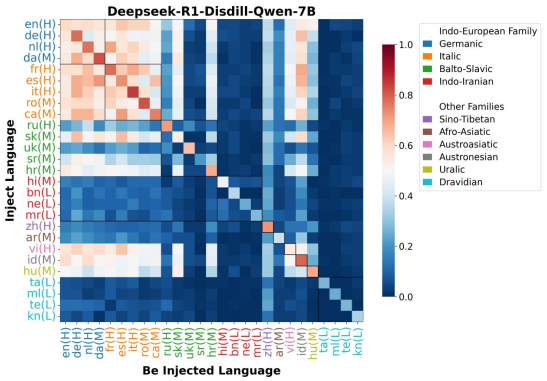

Figure 20: The Influence Degree across languages in Deepseek-R1-Distill-Qwen-7B.

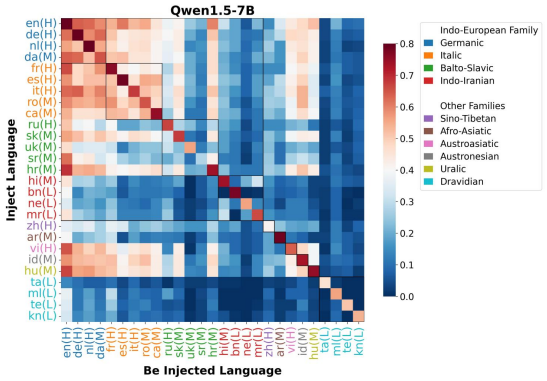

Figure 21: The Editing Success Rate across languages in Qwen1.5-7B.

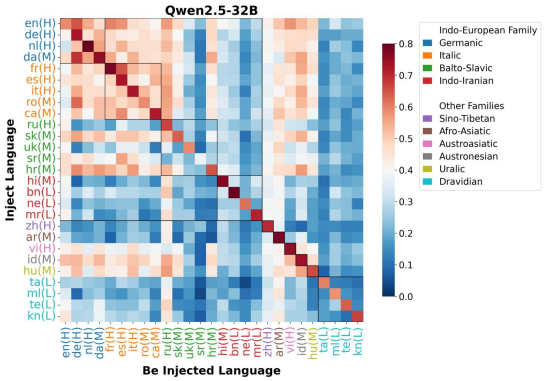

Figure 24: The Editing Success Rate across languages in Qwen2.5-32B.

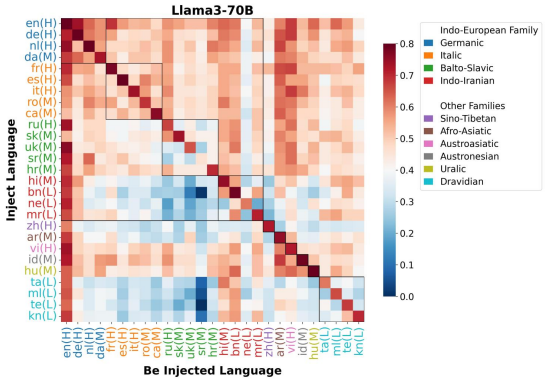

Figure 22: The Editing Success Rate across languages in llama3-70B.

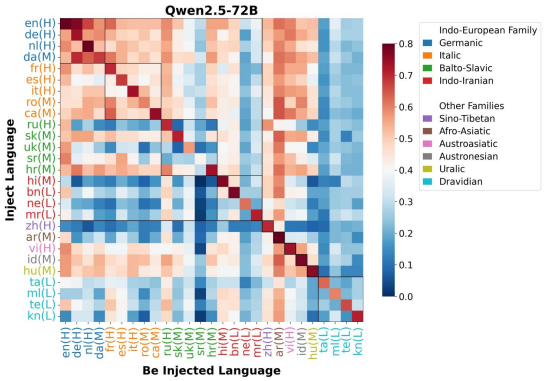

Figure 25: The Editing Success Rate across languages in Qwen2.5-72B.

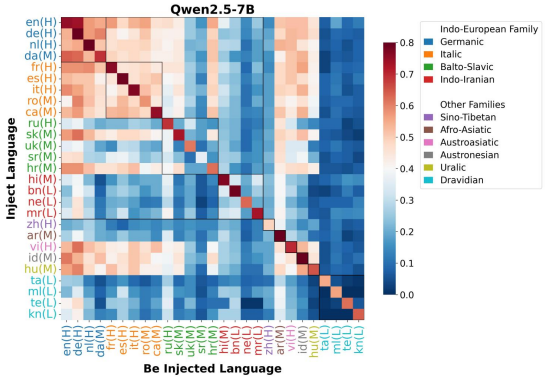

Figure 23: The Editing Success Rate across languages in Qwen2.5-7B.

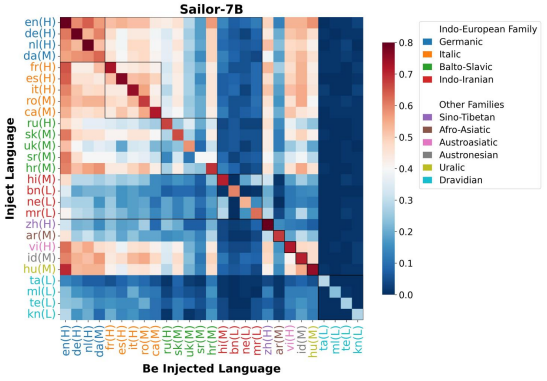

Figure 26: The Editing Success Rate across languages in Sailor-7B.

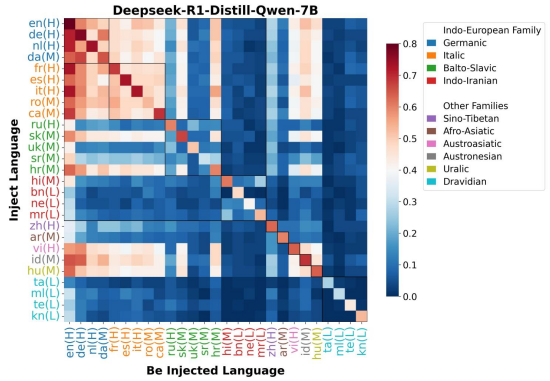

Figure 27: The Editing Success Rate across languages in Deepseek-R1-Distill-Qwen-7B.

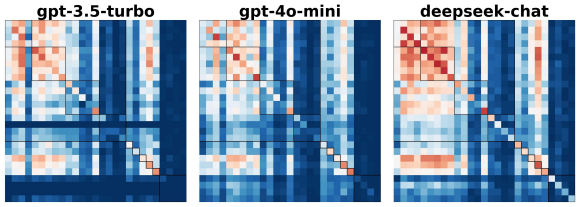

Figure 28: The results of Qwen1.5-7B under three different translators (gpt-4o-mini, gpt3.5-turbo, deepseek-chat) in the Misinformation Propagation setup.

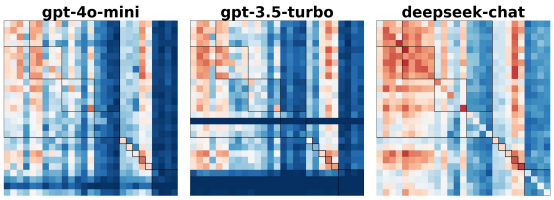

Figure 29: The results of Llama3-8B under three different translators (gpt-4o-mini, gpt3.5-turbo, deepseek-chat) in the Misinformation Propagation setup.

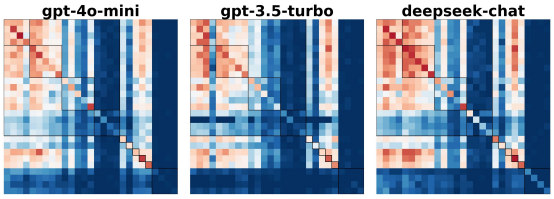

Figure 30: The results of Sailor-7B under three different translators (gpt-4o-mini, gpt3.5-turbo, deepseek-chat) in the Misinformation Propagation setup.