Breaking Boundaries: Investigating the Effects of Model Editing on Cross-linguistic Performance

Somnath Banerjee[†] **Avik Halder**[†*] **Rajarshi Mandal**[†*] **Sayan Layek**[†]
Ian Soboroff[‡] **Rima Hazra**[∓] **Animesh Mukherjee**[†]
[†]Indian Institute of Technology Kharagpur, India
[‡]National Institute of Standards and Technology, USA
[∓]INSAIT, Sofia University "St. Kliment Ohridski"
{som.iitkgpcse}@kgpian.iitkgp.ac.in

Abstract

Pretrained language models (PLMs) have transformed natural language processing (NLP) but tend to exacerbate linguistic disparities in multilingual contexts. While earlier research has primarily focused on transformer-based models like BERT, this study shifts attention to large language models (LLMs) such as MISTRAL, TOWERINSTRUCT, OPENHATHI, TAMIL-LLAMA, and KAN-LLAMA. Through comprehensive evaluations across eight languages—including high-resource ones (English, German, French, Italian, Spanish) and low-resource ones (Hindi, Tamil, Kannada)—the research uncovers significant shortcomings in ensuring multilingual robustness and adaptability. Employing frameworks like "each language for itself" (ELFI) and "each language for others" (ELFO), the analysis reveals that existing LLMs struggle to address linguistic inequities. Even strategies like model merging fail to close these gaps, highlighting fundamental deficiencies. These findings underscore the urgent need to redesign AI systems to achieve genuine linguistic inclusivity and balanced performance across diverse languages.

1 Introduction

Handling multilinguality in language models remains a significant challenge, particularly when models are prompted in languages other than English. Tasks such as question answering (Xu et al., 2024a), addressing multilingual safety concerns (Wang et al., 2024; Deng et al., 2024), or performing knowledge edits (Hazra et al., 2024) often reveal noticeable gaps in performance for low-resource languages. Despite the advancements in multilingual large language models (LLMs), disparities persist, especially for languages with fewer computational resources. A clear example of this issue arises in knowledge editing (Sinitsin et al., 2020; De Cao et al., 2021). For instance, when an LLM is updated to correct a factual statement, "***The PM of the UK is Rishi Sunak***" to "***The PM of the UK is Keir Starmer***" the model may apply the update accurately in well-represented languages like English or French (Qi et al., 2023; Xu et al., 2023). However, the same edit often fails to propagate when queried in low-resourced languages like Tamil or Hindi. This inconsistency highlights a critical weakness in the ability of LLMs to transfer factual updates across languages. Even advanced models like MISTRAL and TOWERINSTRUCT, while effective in European languages, struggle significantly with low-resource languages. This limitation undermines the broader goal of making language technologies universally accessible and equitable (Wang et al., 2023).

This research aims to uncover the disparities in cross-lingual performance of LLMs to promote future linguistic inclusivity. While model editing techniques have advanced in monolingual settings, ensuring that factual updates made in one language are accurately reflected across others remains a major challenge (Hazra et al., 2024; Banerjee et al., 2024). This issue is particularly severe for low-resource languages, where models often fail to maintain reliability and consistency after edits. Such limitations reduce the utility of LLMs for these languages and widen existing linguistic inequities, leaving many communities underserved. Our work highlights these gaps, showing how current models struggle to manage multilingual updates, especially in underrepresented languages. By evaluating cross-lingual performance, we emphasize the need for more inclusive approaches to ensure that LLMs benefit users of all languages, not just those with abundant resources.

In this work, we conduct a comprehensive evaluation of how factual knowledge is transferred and maintained across eight linguistically diverse languages. We examine established knowledge editing techniques such as ROME (Meng et al.,

[*]These authors contributed equally to this work.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 194–209

2022) and MEMIT (Meng et al., 2023) to assess their performance in multilingual contexts. Our research utilizes two strategies (Das et al., 2022)—"*each language for itself*" (**ELFI**) and "*each language for others*" (**ELFO**)—to rigorously test the ability of LLMs to preserve cross-lingual knowledge consistency. Through this evaluation, we reveal current models' limitations in maintaining consistent cross-lingual edits, emphasizing critical gaps to address for enhancing LLMs, particularly in low-resource languages. Our key contributions are as follows.

- We conduct extensive model editing experiments across eight languages—English (**En**), German (**De**), French (**Fr**), Italian (**It**), Spanish (**Es**), Hindi (**Hi**), Tamil (**Ta**), and Kannada (**Kn**)—using **ELFI** and **ELFO**, focusing on decoder-only models' multilingual performance.
- We evaluate 7B decoder-only models, including MISTRAL, TOWERINSTRUCT, OPENHATHI, TAMIL-LLAMA, and KAN-LLAMA, with editing methods **ROME** and **MEMIT**, advancing model editing research.
- This is the first of it's kind work on LLM to reveal that model merging improves capabilities but struggles with cross-lingual consistency after editing.

2 Related work

Targeted parameter editing modifies specific model components to integrate new information. (Dai et al., 2022) introduced adjustments to 'knowledge neurons' in transformers, while ROME (Meng et al., 2022) updated neural weights to refresh LLM knowledge. MEMIT (Meng et al., 2023) expanded ROME for simultaneous updates, with further validation by (Hase et al., 2023; Yao et al., 2023).
Multilingual knowledge editing remains limited, focusing mainly on translating English prompts. X-FACTR (Jiang et al., 2020) and M-LAMA (Kassner et al., 2021) exposed large knowledge gaps in non-English languages, often with $< 10\%$ accuracy. GeoMLAMA (Yin et al., 2022) revealed that native languages may not best access national knowledge. We analyze cross-lingual consistency in multilingual LLMs, extending prior work mostly on BERT (pre LLM era) to diverse LLMs fine-tuned for specific languages (Wang et al., 2023; Beniwal et al., 2024).

3 Task overview

Model editing: Given a language model θ_{pre} and an edit descriptor <kn, a_{new}, a_{old}>, the model editing technique will create an edited model θ_{edit}. So, for an input prompt kn, θ_{pre} has the old prediction a_{old} and after editing θ_{pre}, the edited model θ_{edit} has updated prediction a_{new} without influencing model behaviour on other samples. Thus, given the edit input kn, θ_{pre} does not produce a_{new}; it is θ_{edit} that is designed to produce the output a_{new}.

$$\theta_{edit}(kn) = \begin{cases} a_{new} & \text{if } kn \in I(kn, a_{new}) \\ \theta_{pre}(kn) & \text{if } kn \in O(kn, a_{new}) \end{cases} \quad (1)$$

The scope of consideration, $I(kn, a_{new})$, includes kn and similar versions of it. This means it covers the original input and any rephrased versions of it that still relate to the same topic. For example, if kn is a question, this scope includes different ways of asking the same question. However, the excluded scope, $O(kn, a_{new})$, refers to inputs that are not related to the edit case provided. So, it leaves out any inputs that do not have anything to do with kn or its related versions. Along with the updated information, the edited model should follow the four properties: (i) ***reliability*** – θ_{edit}, produces the correct response for the specific edit scenario represented by (kn, a_{new}), (ii) ***generalization*** – the edited model θ_{edit} must uniformly apply edits to both the designated edit case (kn, a_{new}) and its semantically equivalent variations, guaranteeing a consistent output, a_{new}, across all rephrased iterations of kn, (iii) ***locality*** – θ_{edit} should not alter the output for examples outside its intended scope (O(kn, a_{new})), and (iv) ***portability*** – evaluates the capacity of edited model θ_{edit} for robust generalization, assessed through questions designed to test the edited model's reasoning with updated knowledge.
Multilingual knowledge editing: Given a set of languages $\mathcal{L}$, we consider a language $l \in \mathcal{L}$ to edit the model θ_{pre} and obtain θ^l_{edit}. We then test the edited model θ^l_{edit} with all the languages in $\mathcal{L}$. In the equations below, s is the source language, and t is the target language. The conditions are as follows: if kn_s is in the inclusion scope $I(kn, a_{new})$, the model should output a^s_{new}. Otherwise, if kn_s is in the exclusion scope $O(kn, a_{new})$, the model should output $\theta_{pre}(kn_s)$. For the target language,

similar conditions apply with transformations $\mathcal{T}^t$.

$$\theta_{edit}(kn_s) = \begin{cases} a^s_{new} & \text{if } kn_s \in I(kn, a_{new}) \\ \theta_{pre}(kn_s) & \text{if } kn_s \in O(kn, a_{new}) \end{cases} \quad (2)$$

$$\theta_{edit}(kn_t) = \begin{cases} \mathcal{T}^t(a^s_{new}) & \text{if } kn_t \in \mathcal{T}^t(I(kn, a_{new})) \\ \theta_{pre}(kn_t) & \text{if } kn_t \notin \mathcal{T}^t(O(kn, a_{new})) \end{cases} \quad (3)$$

$\mathcal{T}^t(.)$ transforms the target output of the source language to the target language with the same meaning. Therefore, after editing the model in one language, such as English, the effect of the edit should be reflected in other languages as well. This ensures that the specific edit is consistent across all languages, regardless of the language in which the edit was made.

Model merging: In the specific case of Indic languages – Hindi, Tamil and Kannada – we have specialized LLMs for each unlike in the case of Western languages where the models we have used are known to be pretrained on all those languages. We investigate if the three LLMs for the Indic languages could be further unified to obtain a more powerful model θ_{merged}, which dynamically harnesses the specialized linguistic capabilities of each constituent models. This involves extracting language-specific unique task vectors from instruction-tuned models, i.e., $\theta_{base-Hindi} \rightarrow \vec{v}_{Hindi}$, $\theta_{base-Tamil} \rightarrow \vec{v}_{Tamil}$, and $\theta_{base-Kannada} \rightarrow \vec{v}_{Kannada}$ for each respective language. These vectors are integrated using a TIES (Yadav et al., 2023) merging technique to synthesize θ_{merged}. Subsequently, θ_{merged} is edited in the same process as above to obtain θ_{edit} each time adjusting its output specifically for inputs associated with the defined task and the language.

4 Dataset

For our experiments, we use the popular **CounterFact** (Meng et al., 2022) and **ZsRE** (Levy et al., 2017) datasets. We uniformly sample $\sim$ 550 edit instances from each dataset. Each edit instance in these datasets includes the actual edit case, the reliability prompt, the generalization instances, the locality prompt and its answer, portability and its answer. Further we use google translator [1] to translate each edit instance into seven other languages – German (**De**), French (**Fr**), Italian (**It**), Spanish (**Es**), Hindi (**Hi**), Tamil (**Ta**) and Kannada (**Kn**). In both the datasets, the actual portability prompt is an interrogative sentence (i.e., in the form of question). However, when the question gets translated to other languages, the translated question becomes different from actual question format. For example, when the actual portability prompt in English "To which language family does the official language of Sastamala belong?" is translated to French the new prompt becomes "À quelle langue la famille appartient la langue officielle de Sastamala?". However when this is back-translated to English the prompt means "Which family language does the official language of Sastamala belong to?" which is not the same as the original English prompt. We therefore employed GPT-4[2] to convert question in the interrogative sentence into a task of sentence completion. Subsequently we translate this sentence completion form to other languages to obtain the corresponding portability prompt.

Note to the choice of languages: The Western languages that we choose are based on their cultural, economic and academic significance (Lobachev, 2008)[3] and cover the Romance and the Germanic families. In addition, we include three Indic languages that have far lesser resources compared to their Western counterparts.

5 Experimental setup

5.1 Selection of LLMs

We use the following multilingual LLMs for our experiments:

Mistral-7B-Instruct-v0.2 (MISTRAL)[4]: A multilingual causal language model (Jiang et al., 2023), supporting diverse languages[5].

TowerInstruct-7B-v0.2 (TOWERINSTRUCT)[6]: Based on LLaMA2 (Touvron et al., 2023), supports multilinguality across 10 languages, including English, German, and Chinese.

OpenHathi-7B-Hi-v0.1-Base (OPENHATHI)[7]: Optimized for Indian languages like Hindi and Tamil using a GPT-3-like transformer with hybrid partitioned attention.

Tamil-llama-7b-base-v0.1 (TAMIL-LLAMA)[8]: A bilingual Tamil-English model (Balachandran, 2023) using a 7B-parameter causal language framework.

[1] https://translate.google.com/

[2] openai.com/research/gpt-4, version: gpt-4-0125-preview
[3] https://preply.com/en/blog/most-important-languages/
[4] huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[5] https://encord.com/blog/mistral-large-explained/
[6] huggingface.co/Unbabel/TowerInstruct-7B-v0.2
[7] huggingface.co/sarvamai/OpenHathi-7B-Hi-v0.1-Base
[8] huggingface.co/abhinand/tamil-llama-7b-base-v0.1

Kan-LLaMA-7B-SFT (KAN-LLAMA)[9]: Specialized in Kannada with a 49,420-token vocabulary, pre-trained on 600M tokens from CulturaX using low-rank adaptation. More details on models are in Appendix A.

Languages	Models	TowerInstruct		Mistral		TowerInstruct		Mistral	
		CounterFact				ZsRE			
	Metrics	RO	ME	RO	ME	RO	ME	RO	ME
De	Rel	0.83/0.96	0.73/0.83	0.83/0.96	0.73/0.87	0.48/0.59	0.25/0.30	0.51/0.62	0.38/0.47
	Gen	0.27/0.31	0.19/0.22	0.28/0.31	0.19/0.22	0.33/0.39	0.11/0.12	0.35/0.45	0.18/0.24
	Loc	0.22/0.23	0.19/0.22	0.21/0.23	0.24/0.27	0.00/0.01	0.00/0.01	0.01/0.02	0.01/0.03
	Port	0.01/0.01	0.01/0.01	0.03/0.04	**0.04**/0.06	0.02/0.02	0.00/0.00	**0.08**/0.10	0.02/0.04
Es	Rel	0.82/0.92	0.70/0.80	0.81/0.91	0.78/0.86	0.44/0.59	0.24/0.34	0.49/0.61	0.37/0.49
	Gen	0.33/0.37	0.23/0.27	0.28/0.32	0.22/0.27	0.30/0.40	0.16/0.20	0.35/0.45	0.22/0.29
	Loc	0.21/0.22	0.19/0.19	0.25/0.27	**0.27**/0.29	0.00/0.01	0.01/0.02	0.01/0.01	**0.02**/0.02
	Port	0.00/0.00	0.00/0.00	0.03/0.03	0.03/0.04	0.02/0.02	0.01/0.02	0.03/0.07	0.03/0.04
It	Rel	**0.87**/0.93	0.74/0.78	0.86/0.91	0.80/0.88	0.54/0.62	0.25/0.29	**0.58**/0.65	0.42/0.50
	Gen	**0.35**/0.38	0.25/0.26	0.28/0.30	0.24/0.27	0.35/0.43	0.16/0.20	**0.42**/0.48	0.25/0.31
	Loc	0.18/0.19	0.20/0.20	0.26/0.27	**0.27**/0.28	0.00/0.00	0.00/0.01	0.00/0.02	0.01/0.02
	Port	0.02/0.02	0.02/0.03	0.02/0.03	0.03/0.03	0.01/0.02	0.02/0.03	0.07/0.08	0.01/0.03
Fr	Rel	0.83/0.90	0.65/0.72	0.83/0.89	0.79/0.85	0.51/0.59	0.27/0.35	0.52/0.63	0.40/0.50
	Gen	0.31/0.33	0.22/0.24	0.29/0.30	0.24/0.25	0.28/0.35	0.14/0.17	0.40/0.50	0.19/0.27
	Loc	0.21/0.22	0.17/0.19	0.20/0.22	0.24/0.25	0.00/0.01	0.00/0.02	0.01/0.02	0.01/0.02
	Port	0.00/0.01	0.00/0.00	0.03/0.03	0.03/0.03	0.03/0.05	0.03/0.03	0.06/0.09	0.04/0.06

Table 1: Comparison of reliability, generalization, locality, and portability scores across language models under *Self edit - self inference* settings. The highest scores for individual metrics in ROME and MEMIT are highlighted in magenta for CounterFact and in cyan for ZSRE, with values shown as Exact Match/Partial Match.

Languages	Models	TowerInstruct		Mistral		TowerInstruct		Mistral	
		CounterFact				ZsRE			
	Metrics	RO	ME	RO	ME	RO	ME	RO	ME
De	Rel	0.48/0.53	0.40/0.46	0.50/0.56	0.54/0.61	0.24/0.28	0.10/0.14	0.34/0.45	0.14/0.18
	Gen	0.25/0.27	0.13/0.17	0.23/0.27	0.22/0.23	0.18/0.23	0.12/0.14	0.26/0.35	0.14/0.16
	Loc	0.20/0.21	0.19/0.22	0.23/0.25	0.26/0.28	0.00/0.01	0.00/0.02	0.01/0.02	0.01/0.03
	Port	0.00/0.00	0.00/0.00	0.03/0.03	0.03/0.04	0.02/0.02	0.02/0.02	0.06/0.07	0.02/0.03
Es	Rel	0.51/0.56	0.40/0.48	**0.57**/0.62	0.56/0.60	0.24/0.29	0.12/0.14	**0.39**/0.48	0.19/0.26
	Gen	0.26/0.29	0.18/0.22	0.25/0.29	0.21/0.26	0.18/0.25	0.09/0.11	**0.33**/0.41	0.14/0.21
	Loc	0.22/0.24	0.17/0.17	0.24/0.27	0.25/0.27	0.00/0.01	0.01/0.02	0.01/0.02	**0.02**/0.02
	Port	0.00/0.00	0.00/0.00	0.03/0.03	0.03/0.04	0.02/0.03	0.01/0.01	0.04/0.06	0.04/0.05
It	Rel	0.45/0.50	0.35/0.40	0.47/0.58	0.44/0.49	0.24/0.29	0.12/0.14	0.31/0.34	0.23/0.27
	Gen	0.23/0.27	0.19/0.20	0.25/0.35	0.21/0.23	0.17/0.22	0.11/0.13	0.26/0.32	0.18/0.21
	Loc	0.20/0.21	0.20/0.20	0.24/0.36	**0.28**/0.29	0.00/0.00	0.00/0.01	0.00/0.02	0.01/0.02
	Port	0.01/0.02	0.01/0.02	0.03/0.11	0.04/0.04	0.01/0.02	0.02/0.02	**0.07**/0.08	0.01/0.01
Fr	Rel	0.50/0.53	0.45/0.49	0.49/0.55	0.51/0.59	0.22/0.26	0.12/0.17	0.36/0.44	0.23/0.28
	Gen	**0.28**/0.31	0.19/0.22	**0.28**/0.31	0.26/0.27	0.15/0.21	0.08/0.10	0.29/0.33	0.16/0.21
	Loc	0.23/0.23	0.19/0.21	0.20/0.36	0.25/0.26	0.00/0.01	0.00/0.02	0.01/0.03	0.01/0.02
	Port	0.01/0.01	0.01/0.01	0.01/0.12	0.03/0.04	0.02/0.02	0.02/0.02	0.06/0.09	0.04/0.05

Table 2: Comparison of reliability, generalization, locality, and portability scores across language models under *English edit - self inference* settings. The highest scores for individual metrics in ROME and MEMIT are highlighted in magenta for CounterFact and in cyan for ZSRE, with values shown as Exact Match/Partial Match.

Languages/ Models		*self edit - self inference*				*(English edit - self inference)*			
		CounterFact		ZsRE		CounterFact		ZsRE	
	Metrics	RO	ME	RO	ME	RO	ME	RO	ME
Hi/ OpenHathi	Rel	0.02/0.02	0.45/0.60	0.03/0.06	**0.20**/0.33	**0.56**/0.66	0.02/0.03	**0.03**/0.03	**0.03**/0.06
	Gen	0.00/0.00	**0.26**/0.33	0.01/0.04	**0.19**/0.28	**0.27**/0.34	0.03/0.03	0.03/0.03	**0.04**/0.08
	Loc	**0.31**/0.35	0.02/0.03	**0.01**/0.01	0.00/0.01	**0.26**/0.31	0.03/0.03	**0.00**/0.00	**0.00**/0.01
	Port	**0.01**/0.01	**0.01**/0.01	0.00/0.00	**0.03**/0.03	**0.02**/0.02	0.00/0.01	0.00/0.00	**0.01**/0.01
Ta/ Tamil-Llama	Rel	0.12/0.15	**0.48**/0.59	0.06/0.08	0.16/0.21	0.00/0.00	0.01/0.01	0.00/0.00	0.01/0.01
	Gen	0.03/0.04	0.21/0.25	0.03/0.04	0.10/0.14	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00
	Loc	0.01/0.01	0.01/0.01	0.00/0.00	0.00/0.00	0.01/0.01	0.01/0.02	**0.00**/0.00	**0.00**/0.00
	Port	**0.01**/0.01	**0.01**/0.01	0.00/0.00	0.01/0.01	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00
Kn/ Kan-Llama	Rel	0.21/0.26	0.14/0.18	0.16/0.21	0.05/0.07	0.01/0.01	0.00/0.00	0.00/0.01	0.00/0.01
	Gen	0.07/0.08	0.04/0.05	0.08/0.17	0.05/0.05	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.00
	Loc	0.03/0.04	0.02/0.03	0.00/0.00	0.00/0.00	0.02/0.02	0.03/0.03	**0.00**/0.00	**0.00**/0.00
	Port	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00

Table 3: Comparison of scores in indic language models. Highest scores are in bold, second-highest underlined, with values shown as Exact Match/Partial Match.

5.2 Editing methods

We use ROME (Rank-One Model Editing) (Meng et al., 2022) and MEMIT (Mass Editing Memory in a Transformer) (Meng et al., 2023) which are the state-of-the-art editing schemes and particularly suitable for multilingual settings.

Rank-One Model Editing (ROME): This method specifically alters the weights in the initial feed-forward layers of a pretrained model. It identifies factual associations through causal interventions, enabling precise and effective modifications.

Mass Editing Memory in a Transformer (MEMIT): MEMIT advances ROME, by extending its capabilities. While ROME applied a rank-one modification to the MLP weights of a single layer to embed a memory directly into the model, MEMIT enhances this approach by adjusting the MLP weights across multiple critical layers to incorporate numerous memories.

5.3 Evaluation metric

We evaluate the edited models using two metrics:

Exact match: Here accuracy is determined by checking if the ground truth is present in the model's output. Outputs containing the exact expected response are classified as correct, while others are deemed incorrect, providing a binary measure of performance.

Partial match: The Levenshtein ratio (Levenshtein, 1965) measures textual similarity, calculated as the Levenshtein distance divided by the maximum text length. Outputs surpassing an 80% ratio but not containing the ground truth as a substring are considered accurate, allowing for minor acceptable deviations.

6 Results

6.1 *Self edit - self inference* perspective

In this setup we perform the edit in a particular language (say German) and obtain the generated output from the model in the same language (i.e., German itself).

CounterFact dataset: In our evaluations of the model performance for the **CounterFact** dataset, we observe marked variations across different languages and metrics in Table 1, illustrating significant challenges in multilingual adaptability and contextual understanding. For instance, German language tests show that models like TOWERINSTRUCT and MISTRAL achieve good reliability scores (ROME at 0.83 and MEMIT at 0.73 for TOWERINSTRUCT; the same scores are at 0.83 and 0.73 respectively for MISTRAL). These scores illustrate good model performance in understanding the contextual nuances of German. However, generalization and locality score are less impressive

[9]huggingface.co/Tensoic/Kan-Llama-7B-SFT-v0.5

Dataset		CounterFact								ZsRE							
Inferencing language		En		Hi		Ta		Kn		En		Hi		Ta		Kn	
Editing language	Properties	ROME	MEMIT	ROME	MEMIT	ROME	MEMIT	ROME	MEMIT	ROME	MEMIT	ROME	MEMIT	ROME	MEMIT	ROME	MEMIT
En	Rel	0.73/0.75	**0.95**/0.95	0.00/0.00	0.01/0.01	0.00/0.00	0.01/0.01	0.00/0.01	0.00/0.01	0.29/0.33	**0.59**/0.59	0.01/0.02	0.02/0.02	0.00/0.00	0.00/0.00	0.00/0.02	0.00/0.00
	Gen	0.35/0.35	**0.64**/0.64	0.01/0.01	0.02/0.02	0.01/0.01	0.01/0.02	0.00/0.01	0.00/0.01	0.29/0.31	**0.52**/0.54	0.01/0.02	0.00/0.00	0.01/0.01	0.00/0.00	0.00/0.03	0.00/0.00
	Loc	0.33/0.33	0.27/0.27	0.01/0.01	0.01/0.01	0.02/0.02	0.03/0.03	0.11/0.11	0.12/0.12	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.01/0.01	0.00/0.04	0.01/0.02	**0.02**/0.04
	Port	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.03/0.04	0.02/0.04	0.00/0.01	0.00/0.00	0.00/0.01	0.00/0.00	0.00/0.01	0.00/0.00
Hi	Rel	0.00/0.01	0.01/0.01	0.01/0.03	0.07/0.09	0.00/0.00	0.01/0.01	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.01/0.03	0.05/0.05	0.00/0.00	0.00/0.00	0.00/0.02	0.00/0.01
	Gen	0.00/0.00	0.01/0.01	0.02/0.03	0.03/0.04	0.00/0.00	0.01/0.01	0.00/0.01	0.00/0.01	0.00/0.00	0.01/0.01	0.01/0.03	0.02/0.03	0.01/0.02	0.01/0.02	0.00/0.03	0.00/0.02
	Loc	0.35/0.35	0.35/0.36	0.01/0.01	0.01/0.01	0.03/0.03	0.03/0.03	0.12/0.12	0.13/0.13	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.01/0.01	0.01/0.01	0.01/0.01	0.01/0.01
	Port	0.00/0.00	0.00/0.00	**0.01**/0.01	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	**0.07**/0.08	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.01	0.00/0.01	0.00/0.01	0.00/0.01
Ta	Rel	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.01	0.01/0.01	0.00/0.01	0.00/0.01	0.00/0.00	0.01/0.01	0.00/0.00	0.01/0.01	0.00/0.02	0.01/0.03	0.00/0.01	0.00/0.01
	Gen	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.01/0.01	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.00	0.01/0.01	0.00/0.00	0.01/0.01	0.01/0.01	0.02/0.03	0.00/0.02	0.00/0.02
	Loc	**0.36**/0.36	0.33/0.34	0.01/0.01	0.02/0.02	0.02/0.02	0.02/0.02	0.11/0.11	0.11/0.11	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.01/0.03	0.01/0.02	0.01/0.02	0.01/0.02
	Port	0.00/0.00	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	**0.01**/0.01	0.00/0.01	0.00/0.01	0.00/0.01
Kn	Rel	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.01	0.00/0.00	0.00/0.02	0.00/0.00	0.03/0.03	0.00/0.03
	Gen	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.01	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.01	0.00/0.00	0.01/0.03	0.01/0.02	0.01/0.03	0.00/0.04
	Loc	0.35/0.35	0.34/0.34	0.01/0.01	0.02/0.02	0.03/0.03	0.03/0.03	0.12/0.12	0.12/0.12	0.00/0.00	0.00/0.00	0.01/0.01	0.00/0.00	0.00/0.01	0.00/0.00	0.01/0.01	0.00/0.00
	Port	0.00/0.00	0.00/0.00	0.00/0.01	0.00/0.01	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.01	0.00/0.00	0.00/0.01	0.00/0.00	0.00/0.01	0.00/0.01

Table 4: Comparison of scores across the merged model for three Indic languages, evaluated using the **CounterFact** and **ZsRE** datasets for each language and others. Highest scores are in bold, and second-highest are underlined. Values represent Exact Match/Partial Match results.

(TOWERINSTRUCT at 0.27 and 0.22 on ROME for generalization and locality respectively), indicating difficulties in applying the learned information across broader contexts and different locales within the German language. Similar patterns are observed in Spanish and Italian. In Spanish, TOWERINSTRUCT reaches a reliability score of 0.82 for ROME and 0.70 for MEMIT; for MISTRAL the reliability scores are 0.81 for ROME and 0.78 for MEMIT, suggesting decent grasp of Spanish contexts. However, the generalization scores remain below 0.35 for ROME and locality scores do not exceed 0.29 for MEMIT for any model. Despite TOWERINSTRUCT showing a relatively high reliability in Italian with a ROME at 0.87 and MEMIT at 0.74, the generalization and locality scores remain low (highest being 0.35 on ROME and 0.28 on MEMIT for MISTRAL). In case of the three Indic languages the discrepancies become even more pronounced (See Table 3). OPENHATHI, for example, shows a drastic drop in Hindi, with a ROME reliability of just 0.02 and a MEMIT of 0.45, indicating almost no comprehension of the language nuances. TAMIL-LLAMA and KAN-LLAMA also display low scores across all properties. The highest reliability achieved is 0.21 for ROME for KAN-LLAMA and 0.48 for MEMIT in case of TAMIL-LLAMA, which highlights the limitations in these language models. Portability scores are consistently low across all languages, models, and metrics, demonstrating a significant gap in model training as it fails to effectively account for diverse linguistic structures and cultural contexts.

ZsRE dataset: In case of **ZsRE** dataset (see Table 1) German shows moderate performance in reliability with scores like 0.48 on ROME and 0.25 on MEMIT for TOWERINSTRUCT. The generalization (0.33 for ROME) and locality scores (~ 0) are also very poor. These results indicate substantial deficiencies in capturing language-specific details and generalizing learned information. Spanish fares slightly better in reliability, achieving up to 0.49 on ROME with TOWERINSTRUCT and MISTRAL, but like German, faces challenges in generalization and locality, with the best generality reaching only 0.35 and locality remaining near zero. Italian (It) generally scores higher in reliability, particularly with MISTRAL reaching 0.58 on ROME, though it too struggles with generality and locality. French exhibits a similar trend, with reliability scores reaching up to 0.52 for ROME with MISTRAL and both generalization and locality scores remaining low. Performance markedly drops for the three Indic languages (See Table 3). For instance, Hindi's highest reliability is just 0.03 for ROME, while Tamil and Kannada only achieve maximum reliability scores of 0.06 and 0.16 respectively for ROME. Across all languages, portability scores are low, reflecting limited adaptability and the challenge of transferring learned capabilities from one linguistic context to another.

6.2 *English edit - self inference* perspective

In this setup we perform the edit in a English and obtain the generated output from the model in other languages (e.g., German, Italian etc.).

CounterFact dataset: In German, the reliability scores for models such as TOWERINSTRUCT and MISTRAL suggest moderate effectiveness, with ROME around 0.48 and MEMIT around 0.40 (see Table 2). However, their generalization and locality scores reveal limitations in the models' ability to generalize and localize content effectively with scores not exceeding 0.25 and 0.26 respectively. For Spanish, there is a noticeable improvement in reliability, with ROME scores for MISTRAL

Category	Examples	Possible solution
Lexical ambiguity	English: ‘Fair’ can mean a carnival, treating someone right, or having light skin and/or hair French: ‘Livre’ can refer to a book or to the weight measure pound.	Context-aware models
Syntactic ambiguity	English: “Visiting relatives can be boring.” (Ambiguous: Visiting them, or the relatives who visit, can be boring.) German: “Er sah den Mann mit dem Fernglas.” (He saw the man with the binoculars. Ambiguous: Who has the binoculars?) Italian: “Ho visto l’uomo con il binocolo.” (I saw the man with the binocular. Ambiguous similar to German.)	Better parsing
Semantic ambiguity	French: “Mexx, ça a commencé en” (Mexx, that was started in. Ambiguous: started means founded or started in a particular region) Spanish: “Spike Hughes se origina de” (Spike Hughes originates from. Ambiguous: originates from a place or from a particular family)	Incorporation of additional semantic cues
Cultural ambiguity	English: “Arrow of Time/The Cycle of Time” (Is an album of Peter Michael Hamel. But it could also mean the flow of time) French: “Ce n’est pas ma tasse de thé.” (It’s not my cup of tea. Ambiguous without understanding the idiom.) Italian: “In bocca al lupo.” (In the wolf’s mouth, means good luck. Could be confusing without cultural context.)	Deeper multi-cultural context
Translation errors	English: “In which country’s capital city would you most likely hear Faithless’ original language spoken?” translated into French and back to English becomes “In which country’s capital would you most likely hear the original language of the original spoken”	Reinterpretation of the translation in target language
NER errors	English: “The Little Match Girl” could be a literary fairy tale. Spanish: ‘Rio’ can mean a river or refer to the city Rio de Janeiro.	Integration of knowledge graphs
Idioms	German: “Der Blick von unten” (Literally: Seeing things from a low physical position. Meaning: Considering a situation from a marginalized or disadvantaged perspective.)	Maintain exception lists
Phonetic/orthographic errors	English: ‘Their’ vs. ‘There’ vs. ‘They’re’ Spanish: ‘Vino’ (came) vs. ‘Vino’ (wine)	Context-sensitive correction of word forms
Morphological errors	German: The misuse of gender-specific articles "der" (masculine), "die" (feminine), "das" (neuter) can lead to confusion Italian: Confusion between "mangiato" (eaten) and "mangiando" (eating) can change the temporal context of a sentence.	Integration of specialised morphological rules
Pragmatic errors	French: Using ‘tu’ (informal you) instead of ‘vous’ (formal or plural you) in a formal context can be seen as rude or too casual.	Understanding cultural norms

Table 5: Categorization of multilingual knowledge editing errors, including lexical, syntactic, semantic, cultural, and contextual ambiguities, with examples from English, French, German, Italian, and Spanish, highlighting challenges in cross-lingual consistency and accuracy.

reaching 0.57, and a slight improvement in generalization and locality metrics compared to German. Italian and French show similar trends, with reliability scores peaking at 0.47 for MISTRAL in Italian and 0.49 in French; the generalization and locality scores are still lower. For Tamil and Kannada the reliability are exceptionally low (See Table 3). In fact, in case of Tamil this score is 0 for ROME and 0.01 for MEMIT. Comparatively for Hindi the reliability scores are quite good with 0.56 for ROME. However the portability and generalization scores are again very poor.

Key observations

👉 Models like TOWERINSTRUCT and MISTRAL excel in context-specific reliability but falter in generalization and locality.
👉 Indic languages exhibit larger gaps, reflecting limited linguistic diversity in training.
👉 Cross-lingual edits expose critical weaknesses, with performance dropping across linguistic boundaries, and model merging fails to enhance reliability, locality, or generalization on either dataset.

ZsRE dataset: For languages such as German and Spanish, the models display moderate reliability with MISTRAL, achieving ROME scores up to 0.34 and 0.39 respectively, and MEMIT scores of 0.14 and 0.19 respectively (see Table 2). However, the scores significantly drop for locality and portability, showing that while the models can identify relevant relationships, they struggle to generalize and adapt to the specific linguistic nuances of these languages. The trends are similar in Italian and French, where reliability scores are moderate while locality and generalization scores are poor. Further, for the Indic languages, the score are exceedingly low for all the properties indicating the stark gap in performance highly resource scarce languages.

6.3 *Merged* model perspective

Table 4 presents performance metrics for the merged model, with columns representing inferencing languages and rows indicating editing languages. Editing and inferencing in English yield high reliability scores on the **CounterFact** dataset (ROME: 0.73, MEMIT: 0.95). However, performance drops to near zero when editing in English and inferencing in Hindi, Tamil, or Kannada, exposing the model’s cross-lingual limitations. Editing in Hindi, Tamil, or Kannada consistently results in poor outcomes across all properties, regardless of the inferencing language. This highlights the model’s inability to generalize across linguistic barriers and underscores the need for improved multilingual adaptability. The findings reveal that while the model performs well within the same linguistic environment, its performance deteriorates significantly across lesser-resourced languages, necessitating enhanced training approaches for robust multilingual support.

7 Error analysis

In Table 5 we show the different types of linguistic errors encountered during the translation and editing process. The errors are categorised based on the different types of ambiguities and sheds light on how future models should strengthened by carefully harnessing techniques to tackle these errors. More details are available in Appendix B.

8 Discussion

Here we discuss two important questions – *How do multilingual LLMs handle cross-lingual knowledge edits?* and *What steps can industry practitioners take to address cross-lingual disparities?*

How do multilingual LLMs handle cross-lingual knowledge edits?

Modern *LLMs* often fail to propagate factual updates consistently across languages. While languages like English, French, and German benefit from extensive corpora (Xu et al., 2024b), those like Hindi, Tamil, and Kannada suffer from data scarcity, causing unstable knowledge transfer (Qi et al., 2023). Further, editing methods **ROME** and **MEMIT** encounter problems with highly agglutinative or morphologically rich languages.

Key observations

- **Data scarcity**: Inadequate corpora produce sparse embeddings, disrupting the model's ability to adapt newly introduced facts (Das et al., 2022).
- **Architectural bias**: LLM pipelines typically prioritize English, overlooking morphological idiosyncrasies in languages like Tamil or Kannada.
- **Complex linguistic features**: Idiomatic expressions and cultural references can invalidate edits that were accurate in English (Beniwal et al., 2024); merging specialized models can exacerbate divergences if representations are misaligned (Yadav et al., 2023).

What steps can industry practitioners take to address cross-lingual disparities?

A holistic approach is needed to ensure consistent, multi-lingual fact-editing. Below are five key strategies:

- **Expand low-resource corpora**:
  Rationale: Larger, more representative datasets address embedding sparsity;
  Implementation: Generate crowd-sourced/synthetic data (Hazra et al., 2024).
- **Continuous model editing**:
  Rationale: Iterative edits balance new knowledge with existing facts[a]; primarily important for industries dealing with finance, healthcare, and law (e.g., updating a multilingual LLM to reflect new data privacy laws (GDPR, CPRA) in different regions without retraining from scratch).
  Case study: Microsoft's lifelong editing merges local patches with broader retraining (Cao et al., 2021).
- **Alignment-focused architectures**:
  Rationale: Combine morphological analysis, advanced NER, & cross-lingual parameter sharing;
  Benefit: Stable knowledge propagation in structurally diverse languages (Wang et al., 2023).
- **Dedicated edit modules**:
  Rationale: Log each update & validate in all languages to avoid accidental overwrites;
  Implementation: Use an "edit ledger" in attention layers (Hase et al., 2023).
- **Rigorous multilingual testing:**
  Rationale: Systematic checks prevent bias & misinformation from creeping in;
  Tools: Curated test suites for reliability, cultural fitness, and domain-specific accuracy (Hazra et al., 2024).

[a]https://www.microsoft.com/en-us/research/blog/lifelong-model-editing-in-large-language-models-balancing-low-cost-targeted-edits-and-catastrophic-forgetting/

9 Conclusion

In this study, we investigated the impact of knowledge editing across different languages based on the **CounterFact** and **ZsRE** datasets along with their translations. Our extensive experiments employing a variety of knowledge editing techniques on an array of multilingual LLMs resulted in various crucial observations. We discovered that variations in language-specific model architecture significantly affect the success of knowledge edits, that current editing methods often fail to seamlessly transfer alterations from one language to another, and that modifications made in one language might unexpectedly alter model behavior in another language. This study lays the groundwork for future innovations that could lead to more sophisticated and linguistically inclusive AI technologies.

10 Limitations

Despite the promising results, our study has several limitations. The variability in performance across different languages highlights the inherent challenges in achieving true multilingual consistency, with models exhibiting substantial difficulties in generalizing and localizing edits, particularly in low-resourced languages such as Hindi, Tamil, and Kannada. This discrepancy indicates a need for more inclusive and representative training datasets that encompass a wider range of linguistic and cultural contexts. Additionally, our focus on decoder-only models limits the generalizability of our findings to other types of language models, such as encoder-decoder architectures. The relatively low portability scores across all languages further indicate that current models struggle to transfer learned knowledge effectively from one linguistic context to another, especially in cross-lingual edits where modifications in one language often fail to translate accurately into another. Moreover, the merging of models, while showing some promise, does not consistently improve reliability, locality, or generalization metrics, suggesting that further research is needed to optimize these approaches.

11 Ethical consideration

Our research raises ethical concerns regarding linguistic equity and cultural sensitivity. Disparities in model performance could reinforce existing linguistic inequities, limiting access to AI technologies for speakers of low-resourced languages. Future model development must include diverse languages and dialects to promote equity. Additionally, errors related to cultural ambiguity and idiomatic expressions can lead to misinterpretations or offensive content, necessitating robust evaluation frameworks to ensure cultural sensitivity. Privacy and security risks are also significant, as models may inadvertently reveal sensitive information during knowledge editing processes. Researchers must prioritize user privacy and implement stringent data protection measures to prevent misuse of personal data, ensuring AI technologies are effective and equitable for all users.

12 Potential risk

LLMs can be used for harmful content generation and misinformation spread. The prompts used and generated in this work can be misused to generate harmful content.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *Preprint*, arXiv:2311.05845.

Somnath Banerjee, Sayan Layek, Rima Hazra, and Animesh Mukherjee. 2024. How (un)ethical are instruction-centric responses of llms? unveiling the vulnerabilities of safety guardrails to harmful queries. *CoRR*, abs/2402.15302.

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.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *Preprint*, arXiv:2104.08164.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 32–42, New York, NY, USA. Association for Computing Machinery.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. 2024. Sowing the wind, reaping the whirlwind: The impact of editing language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16227–16239, Bangkok, Thailand. 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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Sergey Lobachev. 2008. Top languages in global information production. *Partnership: The Canadian Journal of Library and Information Practice and Research*, 3(2).

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. *Preprint*, arXiv:2210.07229.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 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.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. In *International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023. Cross-lingual knowledge editing in large language models. *Preprint*, arXiv:2309.08952.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5865–5877, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5554–5569, Toronto, Canada. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024a. A survey on multilingual large language models: Corpora, alignment, and bias. *Preprint*, arXiv:2404.00929.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024b. A survey on multilingual large language models: Corpora, alignment, and bias. *ArXiv*, abs/2404.00929.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Preprint*, arXiv:2306.01708.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu

Zhang. 2023. Editing large language models: Problems, methods, and opportunities. *Preprint*, arXiv:2305.13172.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *Preprint*, arXiv:2205.12247.

A Model selection

Mistral-7B-Instruct-v0.2[10]: The model was developed by (Jiang et al., 2023) and supports multilinguality[11]. It is designed around the causal language modeling framework. We shall refer to this model as MISTRAL.

TowerInstruct-7B-v0.2[12]: This model (Alves et al., 2024) has been developed on top of LLaMA2 (Touvron et al., 2023) architecture and supports multilinguality including English, German, French, Spanish, Chinese, Portuguese, Italian, Russian, Korean, and Dutch. We shall refer to this model as TOWERINSTRUCT.

OpenHathi-7B-Hi-v0.1-Base[13]: The model is designed to optimize multilingual interactions with a special focus on Indian languages. It uses a transformer-based architecture similar to GPT-3 but introduces hybrid partitioned attention to efficiently manage computational resources and enhance responsiveness across languages like Hindi, Tamil, and Bengali. We shall refer to this model as OPENHATHI.

Tamil-llama-7b-base-v0.1[14]: This is a sophisticated model (Balachandran, 2023) developed specifically for bilingual tasks in Tamil and English, leveraging a 7 billion parameter causal language modeling framework. We shall refer to this model as TAMIL-LLAMA.

Kan-LLaMA-7B-SFT[15]: This model is tailored for efficient Kannada text processing with an expanded 49,420-token vocabulary, enhancing its language handling capabilities. Pre-trained on 600 million Kannada tokens from the CulturaX dataset, it employs a low-rank adaptation technique to minimize computational costs while preserving the model's integrity. We shall refer to this model as KAN-LLAMA.

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[11] https://encord.com/blog/mistral-large-explained/

[12] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2

[13] https://huggingface.co/sarvamai/OpenHathi-7B-Hi-v0.1-Base

[14] https://huggingface.co/abhinand/tamil-llama-7b-base-v0.1

[15] https://huggingface.co/Tensoic/Kan-Llama-7B-SFT-v0.5

B Error analysis

Lexical ambiguity Lexical ambiguity occurs when a word has multiple meanings, leading to confusion without context. For instance, the English word "crane" can refer to a bird or construction equipment, a distinction crucial for accurate knowledge representation.

Syntactic ambiguity Syntactic ambiguity arises from sentence structures that can be interpreted in multiple ways. An example is the English sentence "Visiting relatives can be boring," which could imply either the act of visiting relatives is boring or that the relatives being visited are boring. Resolving these ambiguities requires advanced parsing techniques and an understanding of the specific language's syntax to ensure accurate interpretation.

Semantic ambiguity errors Semantic ambiguity pertains to the uncertainty of meaning within a sentence or phrase. For example, "He gave her a ring" could mean a telephone call or presenting a piece of jewelry. Multilingual systems need to discern the intended meaning based on semantic cues and the broader context, a challenging task given the subtlety of cues and cultural specificities in language use.

Cultural and contextual errors These errors occur when language processing fails to account for cultural idioms or context-specific meanings. Phrases like "Piece of cake" in English, meaning something easy, can be misunderstood if taken literally or translated directly into another language without considering idiomatic expressions. Handling these requires deep cultural knowledge and contextual understanding beyond linguistic analysis.

Translation errors Translation errors emerge when converting text from one language to another, often leading to loss of meaning or inaccuracies. These can be particularly problematic in knowledge editing, where precision is paramount. For example, translating idiomatic expressions or culturally specific terms often requires not just a direct translation but a reinterpretation in the target language.

Named entity recognition (NER) errors NER errors involve the incorrect identification or classification of proper nouns in text. For instance, distinguishing between "Rio" as a river or the city of Rio de Janeiro in Spanish requires contextual analysis. Accurate NER is essential for knowledge databases to correctly link information to entities, demanding sophisticated language models that can navigate these nuances.

Idiomatic expression errors Errors in understanding or translating idiomatic expressions can significantly alter the intended meaning. For example, the Italian idiom "Tra il dire e il fare c'è di mezzo il mare" illustrates the difference between saying and doing, a concept that might be lost if translated literally. Addressing these requires an in-depth understanding of both the source and target languages' idioms.

Phonetic and orthographic errors These errors occur with words that sound similar (homophones) or are spelt similarly (homographs) but have different meanings. For instance, "their," "there," and "they're" in English. Multilingual systems must accurately identify and apply the correct form based on context, a challenging task that often requires human-like understanding of language.

Morphological errors Morphological errors refer to the misuse of word forms, affecting the grammatical structure and potentially changing the meaning of sentences. German's gender-specific articles—der, die, das—offer a prime example, where incorrect usage can confuse readers and misrepresent information. Overcoming these demands a robust grasp of linguistic rules and the flexibility to apply them in diverse contexts.

Pragmatic errors Pragmatic errors involve the misuse or misunderstanding of language in social context, such as politeness or formality levels. An example is the inappropriate use of "tu" (informal) and "vous" (formal or plural) in French, which can significantly affect the tone and perceived respectfulness of an interaction. Addressing these requires sensitivity to cultural norms and the social dynamics of language, highlighting the complexity of human communication and the challenges in replicating these nuances in AI systems.

C Hyperparameters

We adopt all essential parameter values from the ROME and MEMIT study for all the LLMs. The details of these hyperparameters are provided in Table 6.

Hyperparameter values	
layers	[5]
fact_token	subject_last
v_num_grad_steps	25
v_lr	5e-1
v_loss_layer	31
v_weight_decay	1e-3
clamp_norm_factor	4
kl_factor	0.0625
mom2_adjustment	false
context_template_length_params	[[5, 10], [10, 10]]
rewrite_module_tmp	model.layers.{}.mlp.down_proj
layer_module_tmp	model.layers.{}
mlp_module_tmp	model.layers.{}.mlp
attn_module_tmp	model.layers.{}.self_attn
ln_f_module	model.norm
lm_head_module	lm_head
model_parallel	true

Table 6: Hyperparameter values (most of the default values extend from ROME and MEMIT setup).

D Worked-out Example

For instance, a model's recognition of "*Dent Island Light, located in: Belgium*" **(Post Edit)** (see Figure 2 should be consistent, irrespective of the language employed. Such consistency is crucial for ensuring a uniform user experience across different languages, thereby democratizing access to information and technology.

E Exact vs partial match

We showcase plot correlations in Figures 2 and 3.

F Romance and Germanic languages

F.1 Language perspective

F.1.1 CounterFact

In case of **CounterFact** dataset, significant disparities are observed in edited model performance across different languages. Edits done with **En** and tested on **En** consistently showed high reliability scores across all models, with MISTRAL achieving nearly perfect reliability at 0.994 and TOWERINSTRUCT at 0.996 (for ROME). However, performances while testing with **De**, **It**, **Fr**, and **Es** were notably lower, particularly in generalisation (in between ∼0.21-0.28 for MISTRAL) and locality (0.20-0.28 for MISTRAL) metrics, indicating challenges in generalization and nuanced information processing in non-English contexts. The portability scores were modest across the board, underscoring a pronounced need for enhanced multilingual model adaptability.

When the edit is conducted with **De** and tested on **De** reliability scores for TOWERINSTRUCT (0.828) and MISTRAL (0.834) (for ROME) are reasonably high indicating strong contextual understanding. However, testing with other languages like **It**, **Fr**,

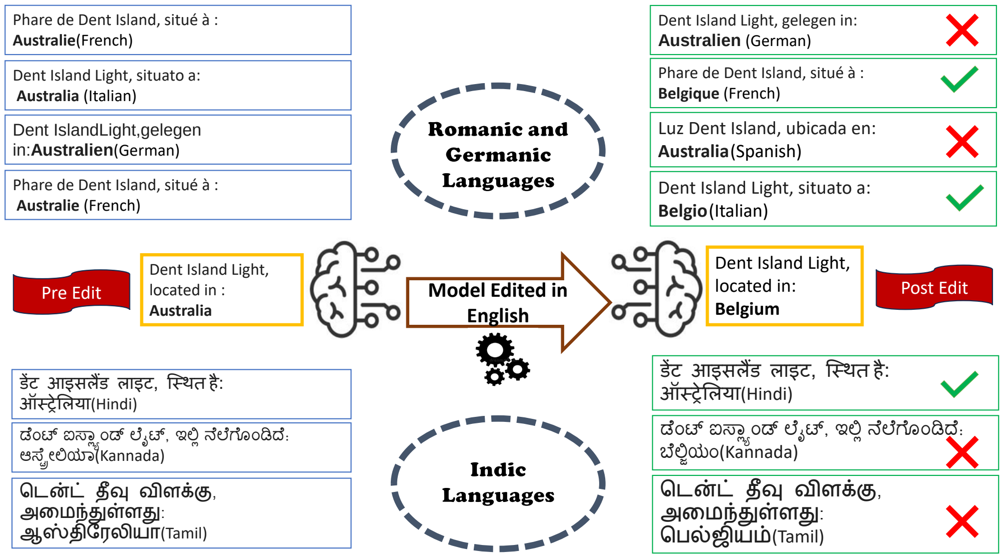

Figure 1: Edited knowledge conflict across various languages for TowerInstruct.

and **Es** exhibit lower scores, reflecting challenges in language-specific processing.

After editing the model with **It** the edited model achieved the highest reliability score with TowerInstruct for test language **It** (0.871) (for ROME). However, the reliability scores for other test languages were lower, with **En** at 0.535, **De** at 0.398, **Fr** at 0.490, and **Es** at 0.488, reflecting the challenge of extending training efficiencies beyond Italian. The highest portability score was seen in **It** with Mistral and TowerInstruct at 0.095 (for ROME), the scores were significantly lower in other languages.

In case of edit with **Fr**, test language **Fr** achieved the highest scores (0.832), with TowerInstruct where it reached 0.454, compared to model's performance in other languages like **En** (0.519), **De** (0.417), **It** (0.509), and **Es** (0.511). This high score in **Fr** for TowerInstruct, however, suggests that certain models can still effectively align with training data even in non-primary languages. In case generality and locality, the scores were universally lower across all models and languages, indicating a struggle in generalizing the **Fr** editing. Locality scores also pointed to difficulties in identifying language-specific nuances, with TowerInstruct showing a modestly better understanding in **It** (0.189) and **Fr** (0.214), yet still remaining low.

After editing with **Es**, **En** (0.555) consistently demonstrated superior reliability score for TowerInstruct, compared to other languages such as **De** (0.391) and **It** (0.451) (excluding **Es**). However, **Es** exhibited notably high reliability scores, with TowerInstruct achieving 0.822 and Mistral 0.812, indicating these models' effective adaptation to Spanish linguistic features. Generality and locality metrics, which measure a model's ability to generalize training and identify language-specific information, respectively, showed universally lower scores across all languages, highlighting challenges in cross-lingual applicability.

F.1.2 ZsRE

After editing with **En** language, the reliability score for Mistral model in **En** was remarkably high at 0.929. However, this contrasts sharply with its performance in other languages such as **De** (0.344) and **It** (0.312), suggesting a significant drop in model effectiveness when transitioning from **En**. Similarly, the TowerInstruct model showed a strong performance when the test langauage was **En** with a relevance score of 0.875, yet scores in other languages like **De** (0.236) and **Fr** (0.221) were markedly lower, highlighting the challenges in maintaining model performance across linguistic boundaries (for ROME). In case of generalization and locality, the scores also emphasize the disparity. While Mistral displayed a good generality in **Eng** (0.812), its scores in languages such as **De** and **It** were only around 0.260. This trend of decreased performance is echoed in the locality scores, where Mistral exhibited almost no ability to identify language-specific nuances in **It** and **Fr**. TowerInstruct's portability score for **En**

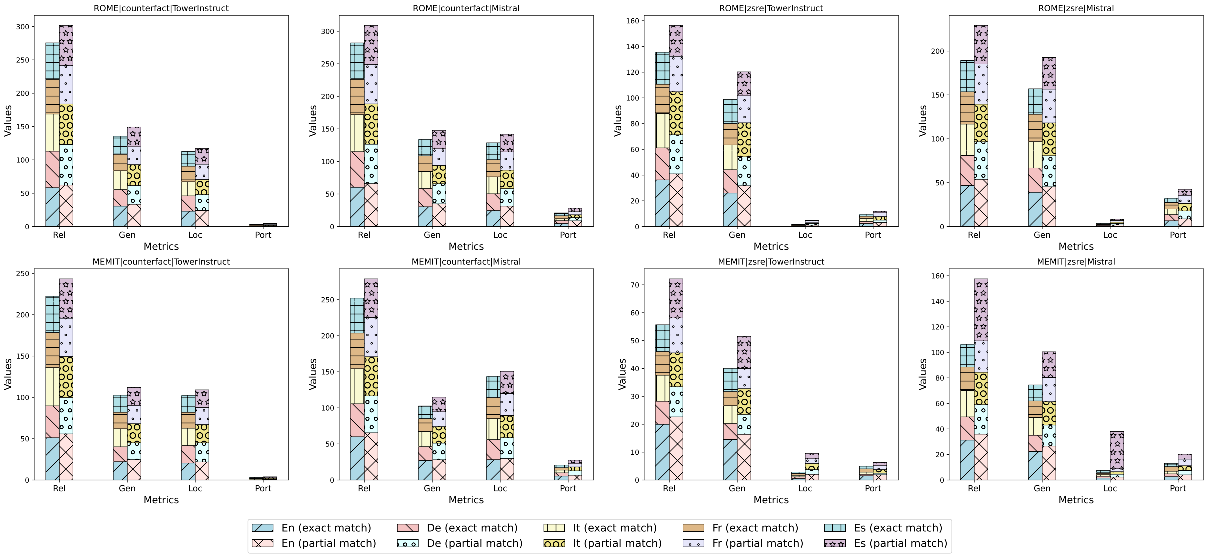

Figure 2: Each metric on the x-axis is represented by two bars: the left bar indicates an exact match, while the right bar indicates a partial match. For each bar, the divisions along the y-axis reflect the average values of the metric, aggregated across Romance and Germanic languages evaluated. These subdivisions are color-coded to denote the editing language, as specified in the legend.

was 0.097, which, although not very high, still outperforms its **De** and **Fr** counterparts, suggesting a somewhat better but still limited ability to adapt training across languages (for ROME).

After editing with **De**, the TOWERINSTRUCT model exhibited significant variations in reliability scores, achieving its highest in **De** (0.480) but only 0.157 in **En**, indicating a substantial challenge in adapting to **De** compared to other languages. Similarly, MISTRAL displayed relatively better relevance in **De** at 0.513, but this still fell short compared to its performance in **It** (0.257), suggesting a consistent trend of models performing better in Romance languages. Further examination of generalization and locality metrics highlights these disparities even more. For instance, generalization scores for MISTRAL in **De** stood at 0.349, yet locality scores were nearly zero across the board, showing a significant deficiency in capturing language-specific details. Portability scores also reflect limited adaptability, with MISTRAL scoring only 0.079 for **De** compared to a slightly better performance in **It** (0.066), underscoring the need for model training approaches that better address and bridge these linguistic gaps to enhance overall performance and applicability across diverse linguistic datasets (for ROME).

After editing with **It**, TOWERINSTRUCT model exhibited a disparity in reliability scores, achieving a high value of 0.537 in **It** but only 0.185 in **De**, underscoring a significant challenge in adapting to **De** compared to other Romance languages. Similarly, MISTRAL demonstrated better reliability in **It** (0.575), further indicating that models tend to align more effectively with training data in certain languages over others. In terms of generality and locality, the scores further emphasize these challenges.

After editing with **Fr**, the TOWERINSTRUCT demonstrated a stronger performance in **Fr** with a reliability score of 0.507 and a generality score of 0.281, compared to its performance in **Es** (Rel: 0.138, Gen: 0.113) and **It** (Rel: 0.197, Gen: 0.167). This indicates a more robust alignment with **Fr** linguistic features. On the other hand, MISTRAL also exhibited its highest reliability in **Fr** (0.517) but struggled in **De** (0.298) and **It** (0.272), further underscoring the varying model efficiencies across languages. These findings highlight significant challenges in model training, where improvements are needed to enhance language-specific understanding and adaptability, ensuring that models perform consistently well across a diverse linguistic spectrum.

After editing with **Es**, TOWERINSTRUCT achieved a high reliability score of 0.443 for **Es**, significantly surpassing its scores in other languages such as **En** (0.232) and **De** (0.148). This trend suggests a stronger model alignment with the linguistic properties of **Es**. In generality, TOWERINSTRUCT

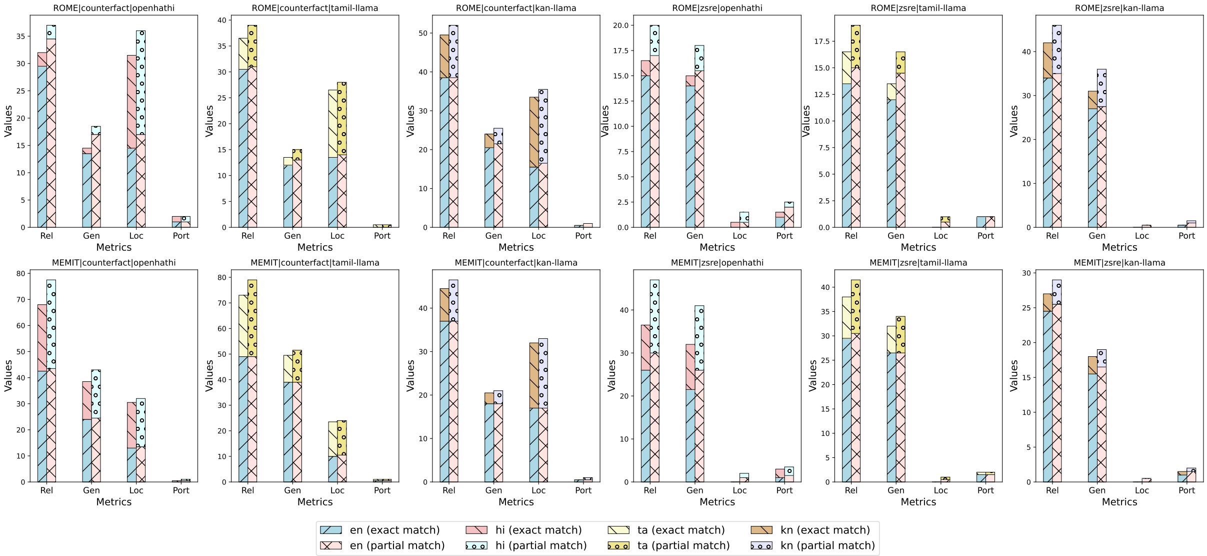

Figure 3: Each metric on the x-axis is represented by two bars: the left bar indicates an exact match, while the right bar indicates a partial match. For each bar, the divisions along the y-axis reflect the average values of the metric, aggregated across all Indic languages evaluated. These subdivisions are color-coded to denote the editing language, as specified in the legend.

highlights better performance in **Es** with a score of 0.305, contrasted with lower scores in **It** (0.202) and **Fr** (0.182). The locality scores were generally low across all languages.

Datasets/ Languages		Score	Mistral					TowerInstruct				
			En	De	It	Fr	Es	En	De	It	Fr	Es
CounterFact	En	Rel	0.994/0.994	0.498/0.560	0.469/0.578	0.487/0.548	0.571/0.617	0.996/0.996	0.482/0.529	0.455/0.500	0.498/0.527	0.511/0.562
		Gen	0.512/0.529	0.233/0.269	0.246/0.346	0.279/0.305	0.252/0.294	0.522/0.538	0.245/0.273	0.231/0.267	0.280/0.309	0.256/0.291
		Loc	0.327/0.338	0.227/0.250	0.240/0.358	0.200/0.362	0.244/0.265	0.307/0.315	0.196/0.207	0.204/0.209	0.225/0.235	0.224/0.238
		Port	0.133/0.144	0.029/0.033	0.027/0.111	0.013/0.119	0.027/0.035	0.005/0.013	0.000/0.004	0.011/0.018	0.005/0.005	0.002/0.004
	De	Rel	0.558/0.591	0.834/0.961	0.471/0.506	0.423/0.471	0.446/0.500	0.589/0.614	0.828/0.959	0.431/0.489	0.439/0.481	0.429/0.497
		Gen	0.355/0.394	0.284/0.313	0.266/0.303	0.255/0.286	0.245/0.282	0.322/0.345	0.271/0.314	0.211/0.246	0.224/0.246	0.224/0.255
		Loc	0.365/0.376	0.208/0.228	0.251/0.264	0.193/0.207	0.263/0.280	0.287/0.292	0.222/0.232	0.212/0.216	0.214/0.224	0.211/0.224
		Port	0.114/0.133	0.029/0.039	0.025/0.027	0.023/0.023	0.033/0.037	0.004/0.014	0.008/0.008	0.004/0.006	0.006/0.012	0.000/0.002
	It	Rel	0.541/0.578	0.422/0.477	0.860/0.914	0.502/0.542	0.519/0.582	0.535/0.564	0.398/0.450	0.871/0.932	0.490/0.535	0.488/0.556
		Gen	0.319/0.346	0.202/0.218	0.278/0.296	0.235/0.239	0.235/0.267	0.330/0.349	0.226/0.253	0.346/0.376	0.263/0.290	0.268/0.311
		Loc	0.350/0.358	0.230/0.251	0.257/0.270	0.210/0.264	0.253/0.265	0.293/0.301	0.199/0.205	0.185/0.189	0.214/0.222	0.203/0.216
		Port	0.095/0.111	0.031/0.045	0.021/0.031	0.012/0.023	0.019/0.031	0.008/0.010	0.004/0.004	0.019/0.021	0.010/0.012	0.006/0.006
	Fr	Rel	0.519/0.548	0.417/0.485	0.509/0.542	0.832/0.890	0.511/0.566	0.530/0.550	0.383/0.440	0.454/0.501	0.827/0.898	0.458/0.506
		Gen	0.282/0.305	0.190/0.215	0.219/0.239	0.294/0.297	0.252/0.268	0.281/0.297	0.200/0.222	0.208/0.230	0.308/0.330	0.234/0.281
		Loc	0.350/0.362	0.243/0.256	0.249/0.264	0.204/0.217	0.276/0.294	0.303/0.316	0.204/0.214	0.189/0.198	0.214/0.220	0.224/0.208
		Port	0.106/0.119	0.020/0.025	0.022/0.023	0.029/0.033	0.023/0.029	0.006/0.018	0.010/0.016	0.010/0.012	0.004/0.006	0.002/0.008
	Es	Rel	0.528/0.548	0.409/0.458	0.483/0.542	0.489/0.544	0.812/0.908	0.555/0.581	0.391/0.429	0.451/0.516	0.466/0.554	0.822/0.921
		Gen	0.297/0.321	0.194/0.217	0.241/0.272	0.231/0.252	0.280/0.315	0.318/0.340	0.184/0.219	0.233/0.251	0.265/0.263	0.330/0.372
		Loc	0.346/0.358	0.235/0.250	0.249/0.262	0.209/0.223	0.254/0.268	0.294/0.300	0.211/0.217	0.186/0.188	0.200/0.238	0.211/0.223
		Port	0.106/0.123	0.022/0.023	0.035/0.037	0.023/0.025	0.029/0.033	0.008/0.014	0.002/0.002	0.008/0.014	0.010/0.020	0.000/0.002

Table 7: Comparison of reliability (Rel), generalization (Gen), locality (Loc), and portability (Port) scores for multiple language models evaluated using the CounterFact dataset and the ROME editing method. The second column indicates the language in which each model was edited.

Datasets/ Languages		Score	Mistral					TowerInstruct				
			En	De	It	Fr	Es	En	De	It	Fr	Es
ZSRE	En	Rel	0.929/0.981	0.344/0.448	0.312/0.344	0.364/0.442	0.390/0.481	0.875/0.928	0.236/0.279	0.240/0.293	0.221/0.260	0.240/0.288
		Gen	0.812/0.851	0.260/0.351	0.260/0.325	0.292/0.331	0.331/0.409	0.620/0.683	0.183/0.226	0.168/0.216	0.149/0.207	0.183/0.255
		Loc	0.000/0.006	0.013/0.019	0.000/0.019	0.013/0.026	0.013/0.019	0.010/0.019	0.000/0.010	0.000/0.005	0.000/0.014	0.005/0.010
		Port	0.097/0.136	0.065/0.071	0.071/0.078	0.058/0.091	0.039/0.058	0.053/0.062	0.019/0.019	0.010/0.024	0.019/0.019	0.019/0.034
	De	Rel	0.382/0.474	0.513/0.625	0.257/0.336	0.289/0.349	0.270/0.355	0.157/0.216	0.480/0.593	0.221/0.260	0.211/0.240	0.176/0.211
		Gen	0.342/0.428	0.349/0.454	0.237/0.309	0.237/0.289	0.217/0.289	0.152/0.196	0.333/0.387	0.162/0.201	0.142/0.172	0.132/0.167
		Loc	0.000/0.007	0.013/0.020	0.000/0.013	0.013/0.020	0.013/0.020	0.010/0.020	0.000/0.010	0.000/0.005	0.000/0.015	0.005/0.010
		Port	0.079/0.092	0.079/0.099	0.066/0.079	0.072/0.099	0.053/0.086	0.010/0.020	0.025/0.025	0.010/0.015	0.020/0.020	0.010/0.015
	It	Rel	0.314/0.386	0.288/0.340	0.575/0.654	0.333/0.399	0.281/0.366	0.176/0.224	0.185/0.215	0.537/0.624	0.210/0.268	0.229/0.340
		Gen	0.340/0.405	0.242/0.281	0.418/0.484	0.294/0.373	0.222/0.327	0.161/0.200	0.137/0.185	0.346/0.429	0.180/0.239	0.122/0.271
		Loc	0.000/0.007	0.013/0.020	0.000/0.020	0.013/0.020	0.013/0.020	0.010/0.015	0.000/0.010	0.000/0.005	0.000/0.015	0.005/0.005
		Port	0.059/0.085	0.072/0.078	0.072/0.085	0.078/0.105	0.039/0.072	0.029/0.029	0.029/0.029	0.015/0.020	0.029/0.034	0.020/0.030
	Fr	Rel	0.424/0.477	0.298/0.344	0.272/0.391	0.517/0.629	0.331/0.444	0.143/0.177	0.153/0.187	0.197/0.256	0.507/0.591	0.138/0.167
		Gen	0.371/0.424	0.285/0.325	0.245/0.325	0.404/0.503	0.245/0.351	0.138/0.177	0.133/0.167	0.167/0.192	0.281/0.350	0.113/0.163
		Loc	0.000/0.007	0.013/0.020	0.000/0.020	0.013/0.020	0.013/0.020	0.010/0.020	0.000/0.010	0.000/0.005	0.005/0.015	0.005/0.010
		Port	0.132/0.159	0.066/0.066	0.073/0.086	0.060/0.093	0.040/0.060	0.015/0.025	0.025/0.025	0.010/0.020	0.034/0.054	0.005/0.020
	Es	Rel	0.367/0.440	0.260/0.320	0.360/0.433	0.307/0.400	0.487/0.607	0.232/0.232	0.148/0.158	0.241/0.340	0.182/0.236	0.443/0.591
		Gen	0.287/0.367	0.227/0.280	0.247/0.313	0.333/0.387	0.353/0.453	0.153/0.177	0.094/0.118	0.202/0.271	0.182/0.241	0.305/0.404
		Loc	0.000/0.007	0.013/0.020	0.000/0.020	0.013/0.020	0.007/0.013	0.010/0.010	0.000/0.005	0.000/0.005	0.000/0.010	0.005/0.010
		Port	0.060/0.080	0.040/0.060	0.033/0.060	0.047/0.080	0.033/0.067	0.000/0.000	0.010/0.010	0.015/0.030	0.010/0.020	0.020/0.020

Table 8: Comparison of reliability (Rel), generalization (Gen), locality (Loc), and portability (Port) scores for multiple language models evaluated using the ZsRE dataset and the ROME editing method. The second column indicates the language in which each model was edited.

Datasets/ Languages		Score	Mistral					TowerInstruct				
			En	De	It	Fr	Es	En	De	It	Fr	Es
CounterFact	En	Rel	0.988/0.988	0.537/0.606	0.438/0.494	0.506/0.588	0.562/0.600	0.954/0.963	0.404/0.459	0.349/0.404	0.450/0.486	0.404/0.477
		Gen	0.444/0.456	0.219/0.225	0.212/0.225	0.263/0.269	0.212/0.263	0.431/0.431	0.128/0.174	0.193/0.202	0.193/0.220	0.183/0.220
		Loc	0.381/0.388	0.256/0.281	0.275/0.287	0.250/0.263	0.250/0.269	0.275/0.294	0.193/0.220	0.202/0.202	0.193/0.211	0.165/0.165
		Port	0.156/0.188	0.025/0.037	0.037/0.037	0.031/0.037	0.025/0.037	0.000/0.000	0.000/0.000	0.009/0.018	0.009/0.009	0.000/0.000
	De	Rel	0.439/0.484	0.726/0.866	0.376/0.420	0.350/0.369	0.363/0.414	0.355/0.391	0.727/0.827	0.282/0.380	0.309/0.309	0.255/0.300
		Gen	0.242/0.280	0.191/0.223	0.185/0.191	0.185/0.217	0.178/0.210	0.227/0.236	0.191/0.218	0.136/0.176	0.182/0.209	0.145/0.164
		Loc	0.376/0.389	0.242/0.268	0.280/0.293	0.229/0.242	0.274/0.280	0.264/0.282	0.191/0.218	0.200/0.231	0.209/0.227	0.200/0.200
		Port	0.108/0.134	0.045/0.064	0.025/0.025	0.013/0.025	0.032/0.051	0.000/0.000	0.009/0.009	0.009/0.009	0.009/0.009	0.000/0.000
	It	Rel	0.372/0.404	0.353/0.410	0.801/0.878	0.455/0.526	0.449/0.526	0.407/0.444	0.361/0.380	0.741/0.778	0.389/0.417	0.426/0.454
		Gen	0.256/0.263	0.141/0.167	0.237/0.269	0.192/0.231	0.179/0.212	0.315/0.315	0.139/0.176	0.250/0.259	0.204/0.213	0.185/0.213
		Loc	0.385/0.397	0.263/0.288	0.269/0.282	0.250/0.263	0.276/0.282	0.269/0.287	0.204/0.231	0.204/0.204	0.194/0.213	0.176/0.176
		Port	0.122/0.147	0.013/0.032	0.026/0.026	0.013/0.019	0.019/0.045	0.009/0.009	0.009/0.009	0.019/0.028	0.009/0.009	0.000/0.000
	Fr	Rel	0.439/0.459	0.395/0.471	0.401/0.433	0.790/0.847	0.446/0.478	0.468/0.477	0.330/0.385	0.330/0.376	0.651/0.716	0.330/0.367
		Gen	0.229/0.268	0.153/0.166	0.159/0.172	0.236/0.255	0.153/0.172	0.294/0.312	0.128/0.147	0.183/0.183	0.220/0.239	0.174/0.193
		Loc	0.389/0.401	0.268/0.293	0.280/0.293	0.242/0.255	0.274/0.280	0.248/0.266	0.183/0.211	0.183/0.183	0.174/0.193	0.174/0.174
		Port	0.089/0.115	0.019/0.032	0.019/0.019	0.025/0.032	0.013/0.025	0.000/0.000	0.009/0.009	0.009/0.018	0.000/0.000	0.000/0.000
	Es	Rel	0.433/0.465	0.338/0.382	0.401/0.452	0.471/0.522	0.777/0.860	0.435/0.463	0.306/0.324	0.370/0.398	0.380/0.398	0.704/0.796
		Gen	0.210/0.229	0.127/0.159	0.121/0.134	0.185/0.217	0.223/0.274	0.241/0.250	0.148/0.157	0.194/0.204	0.213/0.213	0.231/0.269
		Loc	0.395/0.408	0.274/0.306	0.268/0.287	0.242/0.255	0.274/0.287	0.259/0.278	0.194/0.222	0.185/0.185	0.176/0.194	0.185/0.185
		Port	0.108/0.134	0.025/0.051	0.006/0.006	0.013/0.013	0.025/0.045	0.009/0.009	0.000/0.009	0.009/0.019	0.019/0.019	0.000/0.000

Table 9: Comparison of reliability (Rel), generalization (Gen), locality (Loc), and portability (Port) scores for multiple language models evaluated using the CounterFact dataset and the MEMIT editing method. The second column indicates the language in which each model was edited.

Datasets/ Languages		**Score**	**Mistral**					**TowerInstruct**				
			En	**De**	**It**	**Fr**	**Es**	**En**	**De**	**It**	**Fr**	**Es**
ZSRE	**En**	**Rel**	0.786/0.812	0.136/0.182	0.227/0.266	0.227/0.279	0.188/0.260	0.528/0.538	0.104/0.142	0.123/0.142	0.123/0.170	0.123/0.142
		Gen	0.513/0.545	0.136/0.162	0.175/0.208	0.156/0.208	0.136/0.208	0.321/0.330	0.123/0.142	0.113/0.132	0.075/0.104	0.094/0.113
		Loc	0.019/0.026	0.013/0.032	0.013/0.019	0.013/0.019	0.019/0.019	0.019/0.038	0.000/0.019	0.000/0.009	0.000/0.019	0.009/0.019
		Port	0.039/0.065	0.019/0.032	0.006/0.006	0.039/0.052	0.039/0.045	0.019/0.028	0.019/0.019	0.019/0.019	0.019/0.019	0.009/0.009
	De	**Rel**	0.158/0.204	0.382/0.474	0.138/0.178	0.112/0.132	0.118/0.164	0.029/0.077	0.250/0.298	0.048/0.067	0.038/0.058	0.048/0.048
		Gen	0.125/0.171	0.184/0.243	0.138/0.164	0.105/0.118	0.086/0.125	0.058/0.067	0.106/0.115	0.048/0.067	0.038/0.048	0.038/0.058
		Loc	0.020/0.026	0.007/0.026	0.013/0.020	0.013/0.020	0.020/0.020	0.019/0.029	0.000/0.010	0.000/0.010	0.000/0.019	0.010/0.019
		Port	0.039/0.066	0.020/0.039	0.013/0.013	0.007/0.020	0.020/0.033	0.010/0.019	0.000/0.000	0.000/0.000	0.010/0.010	0.000/0.000
	It	**Rel**	0.144/0.176	0.157/0.196	0.425/0.503	0.144/0.183	0.163/0.216	0.019/0.038	0.038/0.067	0.248/0.286	0.067/0.086	0.095/0.124
		Gen	0.105/0.150	0.085/0.118	0.255/0.307	0.144/0.183	0.105/0.157	0.029/0.067	0.048/0.076	0.162/0.200	0.038/0.057	0.048/0.067
		Loc	0.020/0.026	0.007/0.026	0.013/0.020	0.013/0.020	0.020/0.020	0.019/0.029	0.000/0.019	0.000/0.010	0.000/0.029	0.010/0.019
		Port	0.046/0.072	0.007/0.033	0.013/0.033	0.020/0.033	0.020/0.033	0.000/0.010	0.010/0.019	0.019/0.029	0.010/0.010	0.000/0.000
	Fr	**Rel**	0.139/0.172	0.099/0.152	0.166/0.238	0.397/0.497	0.119/0.166	0.048/0.077	0.048/0.067	0.038/0.077	0.269/0.346	0.019/0.058
		Gen	0.152/0.212	0.079/0.139	0.139/0.185	0.185/0.272	0.093/0.139	0.019/0.038	0.029/0.048	0.048/0.077	0.144/0.173	0.010/0.019
		Loc	0.020/0.026	0.013/0.033	0.013/0.020	0.013/0.020	0.020/0.020	0.019/0.029	0.000/0.019	0.000/0.010	0.000/0.019	0.010/0.010
		Port	0.060/0.079	0.020/0.033	0.020/0.020	0.040/0.060	0.040/0.053	0.019/0.019	0.010/0.010	0.000/0.010	0.029/0.029	0.000/0.000
	Es	**Rel**	0.107/0.153	0.073/0.106	0.166/0.213	0.147/0.186	0.373/0.493	0.058/0.087	0.038/0.058	0.087/0.115	0.058/0.106	0.240/0.337
		Gen	0.087/0.256	0.087/0.106	0.140/0.173	0.093/0.146	0.220/0.286	0.048/0.087	0.058/0.087	0.087/0.115	0.058/0.087	0.163/0.202
		Loc	0.020/0.026	0.007/0.026	0.013/0.020	0.013/0.020	0.020/0.020	0.019/0.029	0.000/0.019	0.000/0.010	0.000/0.019	0.010/0.019
		Port	0.033/0.060	0.007/0.013	0.027/0.033	0.033/0.046	0.027/0.040	0.010/0.010	0.000/0.000	0.010/0.010	0.019/0.019	0.010/0.019

Table 10: Comparison of reliability (Rel), generalization (Gen), locality (Loc), and portability (Port) scores for multiple language models evaluated using the ZsRE dataset and the MEMIT editing method. The second column indicates the language in which each model was edited.