Cross-lingual Editing in Multilingual Language Models

Himanshu Beniwal[†*], **Kowsik Nandagopan D**[*], **Mayank Singh**

Discipline of Computer Science and Engineering, Indian Institute of Technology Gandhinagar

{himanshubeniwal, dkowsik, singh.mayank}@iitgn.ac.in

ArXiv Version

Abstract

Our work introduces the **Cross-lingual Model Editing (XME)** paradigm, wherein a fact is edited in one language, and the subsequent update propagation is observed across other languages. To investigate the XME paradigm, we conducted experiments using BLOOM, mBERT, and XLM-RoBERTa on two language families (Latin and Indic). The results reveal notable performance limitations of state-of-the-art Model Editing Techniques (METs) under the XME setting, particularly when the languages involved belong to two distinct families.

1. Introduction

Let us consider updating a language model (in the English language) to reflect the transition of presidential power from Donald Trump to Joe Biden in the United States, using established model editing techniques (Refer to Figure 1). We term this new editing paradigm as **Cross Lingual Model Editing (XME)**. In Figure 2, we refer to the hypernetwork-based editing to illustrate the standard model editing pipeline [2, 1].

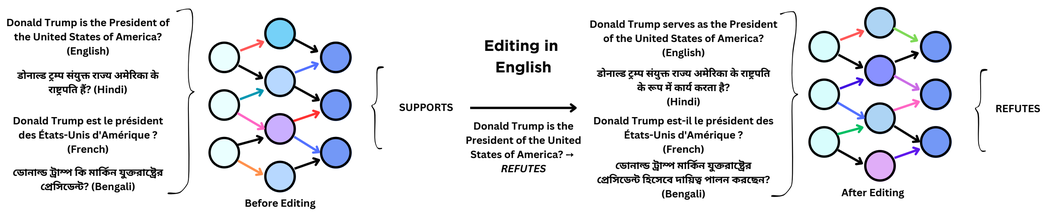

Figure 1: XME pipeline: We update a fact in one language (say English) and check whether the same fact is updated in different languages.

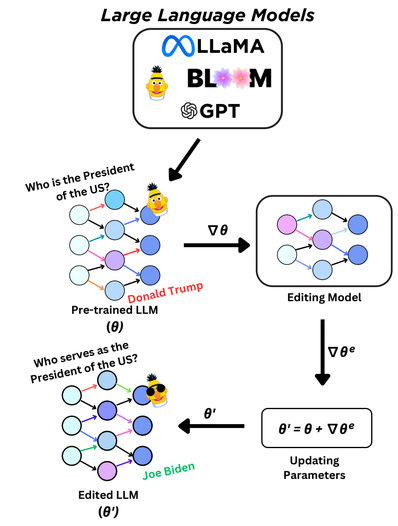

Figure 2: An outline for hypernetwork-based model editing technique.

Research Questions: The primary objective is to address the following research questions:

Q1 *What is the effectiveness of hypernetwork-based editing techniques in cross-lingual settings?*

Q2 *Do different architectures store knowledge at different locations?*

Q3 *How does language selection in the initial fine-tuning stage affect editing performance?*

Q4 *Is the traditional fine-tuning approach more effective than METs in achieving higher performance in the cross-lingual setting?*

Contributions: In our research, we present the following key contributions:

1. We explore the **cross-lingual editing paradigm (XME)** .
2. We uncover a substantial editing performance disparity between **monolingual** and **cross-lingual** contexts.
3. We provide evidence of **distinct knowledge localizations** in different LLMs.

2. Dataset

The statistics for the full multilingual dataset are described in Table 1.

Language	AL_α	AL_β	AL_γ	Train	TFR	VFR
English	11.25	10.67	11.87	104966	10.9998	10.5003
French	10.5	10.6	12.79	104966	10.8479	10.3529
Spanish	12.25	12.53	14.07	104965	10.8479	10.3747
Hindi	14.4	18.04	15.69	103191	10.691	10.2668
Bengali	13.58	20.72	17.61	104966	10.8479	10.3747
Gujarati	15.93	23.86	18.07	104966	10.8479	10.3747
Mixed	11.25	10.67	11.25	102922	10.8633	10.4186
$Inverse_{bloom}$	11.25	-	-	104504	10.8437	10.3747
$Inverse_{xlm}$	-	-	11.95	104966	10.8483	10.3747

Table 1: Dataset statistics in different languages. Note: TFR and VFR are the average length of training-filtered and validation-filtered rephrases, respectively. Inv_{bloom} and Inv_{xlm} are the inverse proportion of BLOOM and XLM-RoBERTa. Lastly, in all the languages, the size of validation and test remains 10444 and 1193, respectively.

3. Evaluation

The Model-Editing techniques are evaluated using two metrics as described below:
Generalizability Score (G_S) assesses the ability of the MET to predict updated facts on semantically equivalent inputs accurately.
Specificity Score (S_S) evaluates the MET's ability to avoid updating unrelated information. In this context, we define an unrelated input as $\hat{x}$, where $\hat{x}$ is irrelevant to the editing fact x.

4. Results

Results are showcased in Table 2 and Table 3, which shows the performance measured by G_S and S_S, respectively. Additionally, Figure 3 and 4 show the best fine-tuning language setting in *monolingual*, *mixed*, and *inverse* proportions.

[†]This work is supported by the Prime Minister Research Fellowship.

[*]Equal Contribution.

		$G_S(x') \rightarrow$						$G_S(x') \rightarrow$					
Set	$x \downarrow$	**en**	**fr**	**es**	**hi**	**gu**	**bn**	**en**	**fr**	**es**	**hi**	**gu**	**bn**
IL	en	**91.79**	87.51	87.85	58.93	52.56	55.24	**87.93**	79.8	80.72	59.93	48.37	58.26
	fr	90.86	**96.9**	92.54	58.59	51.89	55.83	76.36	**87.43**	81.81	58.26	49.29	56.92
	es	90.19	91.79	**95.22**	59.09	52.72	55.99	77.03	80.81	**87.68**	59.51	48.37	56.16
	hi	57.25	58.59	59.68	**96.31**	63.7	71.84	50.88	52.89	52.98	**65.8**	48.7	58.26
	gu	52.64	52.22	53.65	70.41	**95.22**	73.68	50.46	51.63	51.97	53.06	51.47	**57.59**
	bn	54.15	54.06	55.24	71.33	66.14	**96.65**	49.96	51.8	51.55	53.56	49.04	**65.55**
ML	en	**96.56**	94.13	94.97	75.44	62.95	72.09	**93.04**	90.7	88.77	65.55	54.99	69.32
	fr	91.79	**97.99**	96.14	72.34	62.7	69.66	86.17	**89.69**	88.27	64.46	54.57	66.97
	es	90.44	94.72	**97.65**	72.51	62.61	70.33	85.41	**89.44**	89.1	64.21	54.82	65.72
	hi	59.85	63.29	65.21	**96.9**	86.5	87.76	55.41	59.35	58.26	74.1	70.16	**75.27**
	gu	53.48	54.23	56.41	82.31	**96.14**	89.27	55.49	57.75	56.92	73.6	62.7	**76.61**
	bn	55.66	57.59	59.43	82.4	86.92	**97.15**	53.9	56.66	55.57	72.42	**73.26**	71.08
LL	en	**99.67**	99.08	99.25	71.33	59.93	64.04	**85.83**	78.79	79.97	58.09	48.53	63.2
	fr	88.43	**99.83**	98.91	69.91	58.09	63.37	65.97	**89.19**	78.21	59.26	48.7	64.46
	es	75.94	90.78	**94.64**	62.87	57.17	59.18	64.46	74.94	**87.26**	60.86	49.04	66.55
	hi	59.26	75.78	77.87	**100.0**	90.36	91.45	53.06	53.48	**53.9**	43.59	48.45	49.2
	gu	53.06	58.42	66.22	85.5	**99.16**	90.11	51.21	**53.14**	52.98	50.71	50.29	45.52
	bn	56.08	65.72	68.82	90.53	94.22	**99.67**	52.72	**54.15**	53.4	46.19	47.86	47.53
RL	en	**91.79**	84.07	86.84	65.13	55.74	63.54	**88.94**	85.83	85.75	54.32	51.05	62.95
	fr	86.76	**93.21**	86.92	59.01	53.56	57.5	82.31	**88.35**	85.16	53.4	52.64	61.44
	es	86.34	83.24	**92.46**	59.43	53.48	56.83	80.97	82.73	**87.85**	53.06	53.56	61.27
	hi	58.84	56.08	57.33	**92.2**	64.8	68.57	53.81	**56.75**	56.5	51.72	52.98	51.89
	gu	53.4	52.56	53.4	68.15	**92.2**	71.84	54.15	**56.92**	56.33	54.23	32.86	45.1
	bn	55.66	53.56	54.99	67.14	66.3	**92.79**	53.81	**56.08**	55.91	41.99	45.77	37.8

Table 2: The table represents G_S for fine-tuned mBERT (left) and BLOOM (right) on '**en**' dataset using **MEND**.

		$S_S(\hat{x}) \rightarrow$						$S_S(\hat{x}) \rightarrow$					
Set	$x \downarrow$	**en**	**fr**	**es**	**hi**	**gu**	**bn**	**en**	**fr**	**es**	**hi**	**gu**	**bn**
IL	en	98.32	98.09	**98.41**	97.76	98.2	97.48	82.52	93.23	91.37	99.06	99.08	**99.1**
	fr	**98.76**	97.72	98.43	98.26	98.45	97.92	86.8	86.61	92.52	99.62	99.64	**99.73**
	es	**98.58**	98.07	98.16	98.24	98.51	97.76	86.44	93.57	88.68	**99.67**	99.64	99.62
	hi	**98.99**	98.55	98.97	95.03	97.42	96.81	87.49	96.52	94.17	99.56	**99.92**	99.85
	gu	98.89	98.78	**98.99**	96.17	91.49	95.18	87.09	96.4	94.13	**99.85**	84.79	99.83
	bn	98.95	98.62	**99.04**	96.71	96.63	93.0	87.74	96.42	94.3	**99.85**	99.77	97.42
ML	en	97.61	96.69	97.13	97.65	**98.01**	97.11	73.55	83.53	83.45	96.84	**96.94**	**96.94**
	fr	**97.97**	96.23	97.38	97.84	97.95	96.92	82.0	84.74	86.69	97.99	98.01	**98.11**
	es	**98.2**	96.94	96.48	97.65	97.8	97.11	80.68	86.67	83.93	98.53	**98.55**	98.53
	hi	**98.89**	98.41	98.45	91.76	90.82	92.6	93.61	96.33	94.78	99.25	**99.67**	99.22
	gu	**99.02**	98.66	98.74	93.46	83.97	91.34	92.77	96.88	95.03	**99.71**	93.38	98.99
	bn	**98.91**	98.41	98.51	93.67	91.64	88.77	92.77	96.35	94.97	**99.67**	99.62	96.5
LL	en	**99.18**	98.39	98.28	98.81	98.58	98.72	71.94	90.4	89.0	**97.46**	97.4	**97.46**
	fr	**99.45**	92.62	98.01	98.28	99.1	98.07	91.64	92.88	95.16	99.81	99.83	**99.87**
	es	**99.35**	98.11	96.08	98.13	98.64	97.97	91.97	95.2	93.08	99.73	**99.77**	**99.77**
	hi	**99.37**	97.82	97.88	79.59	88.27	87.22	96.33	97.02	95.98	99.43	99.6	**99.62**
	gu	**99.52**	98.32	97.44	90.51	69.32	88.54	96.63	97.23	96.17	**99.77**	94.51	99.45
	bn	**99.33**	97.88	97.74	88.27	86.73	71.86	96.58	97.11	96.81	**99.79**	98.99	97.17
RL	en	97.74	97.02	97.4	97.46	**98.37**	97.53	78.27	88.12	89.12	97.36	97.4	**97.48**
	fr	98.43	95.62	97.32	97.76	**98.64**	97.57	84.62	71.86	77.26	**96.88**	96.67	95.85
	es	**98.34**	97.46	96.65	97.72	98.2	97.65	86.21	77.91	79.15	97.74	**97.8**	97.48
	hi	**98.62**	98.01	98.18	93.94	96.0	94.87	93.9	92.88	92.94	99.75	**99.92**	99.83
	gu	**98.76**	98.51	98.45	95.28	92.71	94.32	94.19	93.8	93.71	**99.96**	96.31	99.77
	bn	**98.72**	98.32	97.99	95.31	95.98	93.11	94.09	92.08	92.44	**99.89**	99.87	98.26

Table 3: The table represents S_S for fine-tuned mBERT on the '**en**' (left) and '**hi**' (right) dataset using **MEND**.

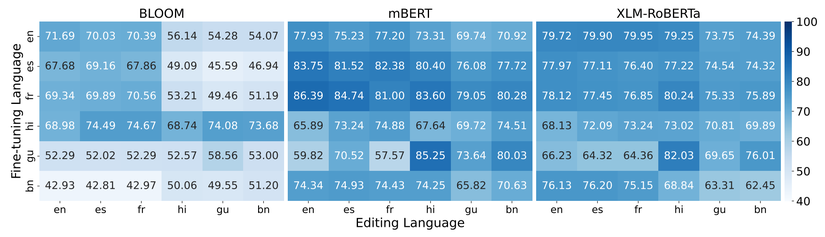

Figure 3: The figure illustrates G_S given the editing language (x-axis) and fine-tuning languages (y-axis) for all the three models BLOOM (left), mBERT (middle) and XLM-RoBERTa (right) when edited using **MEND**.

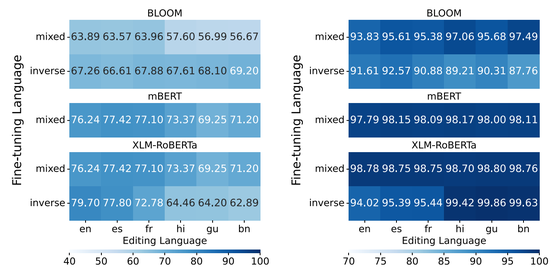

Figure 4: The figure illustrates G_S (Left) and S_S (Right) given the editing language (x-axis) and fine-tuning datasets (y-axis) for all the three models BLOOM (top), mBERT (middle) and XLM-RoBERTa (right) when edited using **MEND**.

Future Directions

We further plan with: ***(1)*** Encoder-Decoder architectures, ***(2)*** extending to more language families, ***(3)*** extending to other NLP Tasks (modeling and translations), and ***(4)*** automating the selection of editing language and layers by modifying the hypernetworks.

Conclusion

Using two distinct language families (six + two language configurations) as our experimental basis, we highlight the storage patterns of factual associations in encoder-only and decoder-only models (three models) with three editing techniques over four different sets of layers.

References

[1] Nicola De Cao, Wilker Aziz, and Ivan Titov. 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, November 2021. Association for Computational Linguistics.

[2] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022.