An Exploration of Knowledge Editing for Arabic

Basel Mousi **Nadir Durrani** **Fahim Dalvi**
Qatar Computing Research Institute, HBKU, Doha, Qatar
{bmousi,ndurrani,faimaduddin}@hbku.edu.qa

Abstract

While Knowledge Editing (KE) has been widely explored in English, its behavior in morphologically rich languages like Arabic remains underexamined. In this work, we present the first study of Arabic KE. We evaluate four methods (ROME, MEMIT, ICE, and LTE) on Arabic translations of the ZsRE and Counterfact benchmarks, analyzing both multilingual and cross-lingual settings. Our experiments on Llama-2-7B-chat show show that parameter-based methods struggle with cross-lingual generalization, while instruction-tuned methods perform more robustly. We extend Learning-To-Edit (LTE) to a multilingual setting and show that joint Arabic-English training improves both editability and transfer. We release Arabic KE benchmarks and multilingual training for LTE data to support future research.

1 Introduction

Despite their impressive capabilities, LLMs suffer from a fundamental limitation: **their knowledge is static and cannot be easily updated without costly retraining or model re-deployment.** This becomes particularly problematic when models must adapt to new facts or correct outdated or incorrect information. To address this, the field of *Knowledge Editing (KE)* has emerged, offering techniques to surgically modify specific factual content within an LLM without retraining from scratch (Wang et al., 2024b; Yao et al., 2023).

Recently, multilingual knowledge editing has garnered some attention (Tamayo et al., 2024; Si et al., 2024; Zhang et al., 2025; Wu et al., 2025; Xu et al., 2023; Durrani et al., 2025). However, the progress on Arabic remains notably limited. **Arabic NLP** poses unique challenges due to diglossia, rich morphology, and the lack of curated resources (Habash et al., 2024; Guellil et al., 2021; Sawaf et al., 2023). The absence of Arabic-specific knowledge editing benchmarks and evaluations creates a significant barrier to understanding how existing KE methods perform in this context.

Furthermore, in today's multilingual world, updating knowledge in one language should ideally generalize to others. This raises critical questions around *multilingual and cross-lingual knowledge editing*: i) Can an edit made in Arabic propagate cross-lingually? ii) Do the same methods perform equally across languages? iii) How can models be trained to edit themselves effectively in multiple languages?

In this work, we present the first study of knowledge editing in Arabic. We benchmark four methods (ROME, MEMIT, ICE, and LTE) on Arabic translations of the ZsRE and Counterfact datasets, evaluating their performance in both multilingual and crosslingual settings.

A central contribution of our work is **extending the Learning to Edit (LTE) framework** to support Arabic and joint Arabic and English training. This multilingual extension improves both editability and crosslingual generalization, demonstrating that instruction-tuned models can adapt edits across languages. We find that parameter-based methods perform inconsistently across languages and exhibit poor transfer. In contrast, LTE delivers strong performance in both Arabic and crosslingual scenarios. To support future research, we release our datasets and multilingual LTE training resources.

Our contributions:

- We analyze four KE methods (ROME, MEMIT, ICE, and LTE) on Arabic edits.
- We compare editing effectiveness across Arabic, English, and German.
- We extend LTE to multilingual settings and evaluate its crosslingual impact.
- We release Arabic versions of ZsRE and Counterfact for KE evaluation.
- We provide multilingual training data for instruction tuned editing.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 417–424

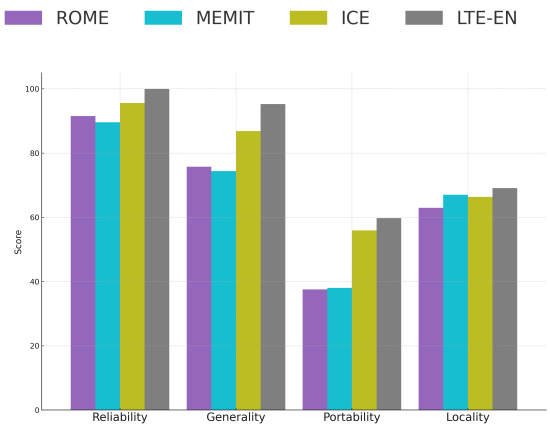


Figure 1: Comparison of ROME, MEMIT, and ICE on LLaMA2-7B-Chat across four metrics: reliability, generality, locality, and Portability

2 Preliminaries

Knowledge Editing (KE) updates a language model f_θ with a new fact (x_e, y_e), producing an edited model f_{θ_e} that satisfies $f_{\theta_e}(x_e) = y_e$ while preserving unrelated outputs.

We evaluate KE using four standard metrics: **reliability** (accuracy on the edit), **generality** (consistency on paraphrases), **locality** (preservation of unrelated knowledge), and **portability** (reasoning with the edited fact in new contexts).

In the **multilingual setting**, edits and evaluations occur within the same language ℓ. In the **cross-lingual setting**, edits are applied in one language ℓ_i and evaluated in another ℓ_k.

3 Experimental Setup

3.1 Data Curation

To enable knowledge editing research in underrepresented languages, we construct **Arabic and German versions of two widely used KE benchmarks**: *ZsRE* and *Counterfact*.

ZsRE (Levy et al., 2017) was originally introduced for zero-shot relation extraction and later adapted for KE by (De Cao et al., 2021; Mitchell et al., 2022). It consists of well-defined factual triples and serves as a strong basis for evaluating *reliability* and *generality* in KE.

Counterfact (Meng et al., 2022a) was designed to test model robustness under *counterfactual* knowledge-false facts that plausibly contradict known information. This benchmark is especially useful for evaluating *locality*, i.e., ensuring that edits do not bleed into unrelated knowledge.

Translation and Release. We use the NLLB-200 model[1] (Team et al., 2022) to automatically translate ZsRE and Counterfact into Arabic and German. While synthetic, these translations are high-quality and provide the first large-scale KE benchmark for Arabic. [2]

Our Contribution. Several datasets were developed for multingual knowledge editing (Wei et al., 2025; Wang et al., 2024c,a; Wu et al., 2023; Nie et al., 2025; Ali et al., 2025). To the best of our knowledge, this is the **first release of Arabic knowledge editing benchmarks** based on ZsRE and Counterfact. Each sample is aligned with evaluation protocols for *reliability*, *generality*, *locality*, and *portability*, making the data immediately usable for reproducible multilingual KE research.

We use the standardized splits from the KnowEdit benchmark (Zhang et al., 2024) and preserve their structure to ensure compatibility with prior work.

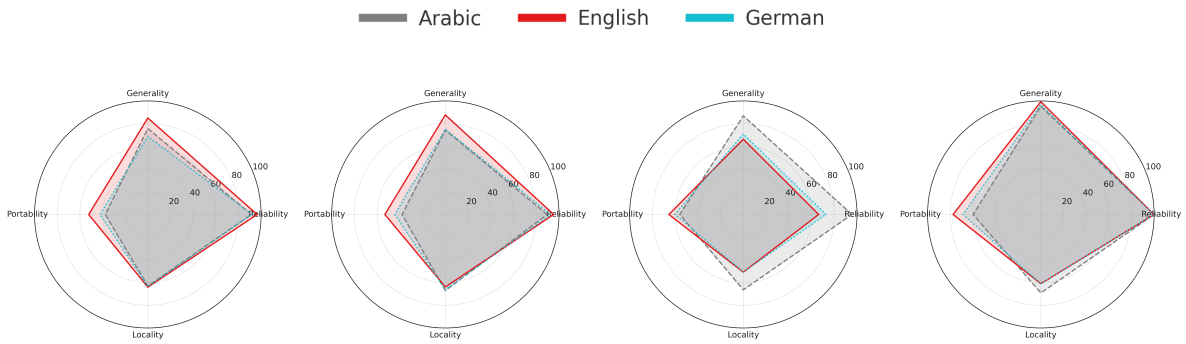
3.2 Knowledge Editing Methods

To evaluate knowledge editing in Arabic and cross-lingual contexts, we compare four representative methods spanning distinct paradigms: **ROME** (Meng et al., 2022a) and **MEMIT** (Meng et al., 2022b) (parameter-based), **ICE** (Zheng et al., 2023) (in-context), and **LTE** (Jiang et al., 2024) (instruction-tuning). While the first three offer complementary approaches to editing and generalization, our primary focus is on extending LTE, given its flexibility and potential for multilingual adaptation.

Originally designed for English, **LTE** fine-tunes models to follow edit instructions through supervised examples, enabling edits to be applied on-the-fly via prompting. We build on this framework by developing both monolingual (Arabic-only) and bilingual (Arabic+English) variants, aiming to assess how instruction diversity impacts editability in Arabic and the model's ability to generalize across languages. This extension allows us to investigate whether LLMs can learn to edit themselves across linguistic boundaries, highlighting the promise of LTE as a foundation for scalable, instruction-driven multilingual editing.

[1] https://hf.co/facebook/nllb-200-3.3B

[2] https://github.com/baselmousi/arabic-knowledge-editing



(a) ROME - ZsRE (b) MEMIT - ZsRE (c) ICE - ZsRE (d) LTE-EN - ZsRE

Figure 2: Impact of the editing language on the reliability, generality, portability and locality metrics on the ZsRE and Counterfact datasets for Llama2-7B-Chat

4 Results and Analysis

4.1 Arabic Editing Performance

How effective are existing knowledge editing methods when applied to Arabic? Figure 1 compares four editing methods: *ROME*, *MEMIT*, *ICE*, and *LTE-EN* on Arabic edits using ZsRE dataset (Counterfact results are shown in figure 5 in Appendix A). **LTE-EN** consistently achieves the highest scores across reliability, generality, locality, and portability, indicating that instruction-tuned models, even when trained only on English, can generalize effectively to Arabic. **ICE** ranks second in reliability and generality, though its portability drops sharply on Counterfact, likely due to the challenge of counterfactual reasoning under zero-shot prompts. **MEMIT** excels in locality, preserving unrelated knowledge via its surgical update mechanism, but trails in generality and portability. **ROME** performs worst overall, highlighting the difficulty of transferring localized parameter edits to morphologically rich, non-English languages.

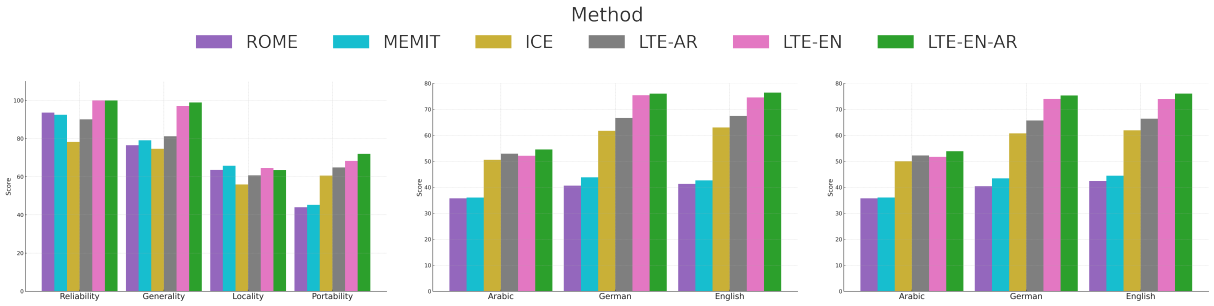
4.2 Multilingual Comparison

LLMs encode different languages in partially overlapping latent spaces (Mousi et al., 2024). This raises an important research question: **How does editing in Arabic compare to editing in other languages?**

To assess cross-lingual robustness, we compare editing performance in **Arabic, English, and German** across four methods: *ROME*, *MEMIT*, *ICE*, and *LTE-EN* as shown in Figure 2 (Counterfact results are shown in figure 6). **Parameter-based methods** (*ROME* and *MEMIT*) perform best in English but degrade noticeably in German and further in Arabic, reflecting their limited adaptability beyond English-tuned settings. In contrast, **ICE** exhibits stable performance across all three languages (Figure 2c), suggesting that prompt-based approaches are more resilient to linguistic variation. Similarly, **LTE** shows minimal degradation across languages, highlighting the benefits of instruction tuning for multilingual generalization.

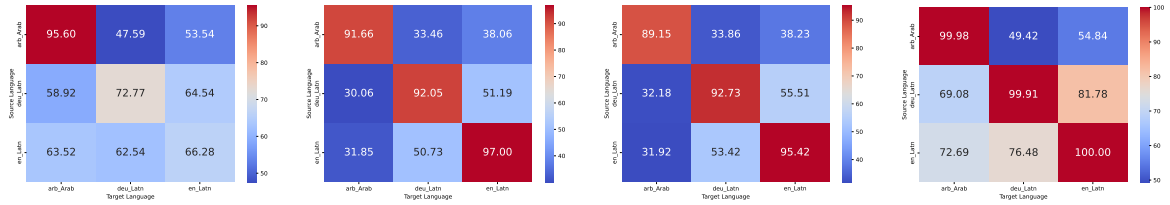
4.3 Cross-Lingual Transfer and Anisotropy

Does editing a fact in Arabic propagate effectively to other languages, and vice versa? A core objective of multilingual knowledge editing is enabling factual edits to transfer seamlessly across languages (Wang et al., 2024a; Khandelwal et al., 2024; Beniwal et al., 2024). To test this, we evaluate bidirectional transfer performance between Arabic, English, and German using the ZsRE benchmark. We consider two setups: (a) editing in Arabic and evaluating in other languages and (b) editing in English or German and evaluating in Arabic. Figure 4 reports the reliability metric on the ZsRE dataset (Appendix A contains additional results on the counterfact dataset). **We observe a clear asymmetry in cross-lingual transfer:** edits made in Arabic fail to propagate reliably to English or German, and vice versa. Parameter-based methods such as ROME and MEMIT show especially weak transfer, confirming that their internal representations are language-sensitive and fail to support consistent multilingual alignment. Even when editing semantically equivalent facts across languages, the models do not generalize edits effectively without explicit multilingual support.



(a) Multilingual Performance of Methods (b) Cross-Lingual Reliability (c) Cross-Lingual Generality

Figure 3: (a) Shows a comparison of the considered methods across the *reliability, generality, locality, and portability* metrics on the ZsRE dataset. (b) Shows a comparison of the averaged cross-lingual reliability scores on the ZsRE dataset and (c) Shows a comparison of the averaged cross-lingual generality scores on the ZsRE dataset. The x-axis in (b) and (c) refer to the language the edit is being applied in.



(a) ICE (b) Rome (c) Memit (d) LTE-EN

Figure 4: Cross Lingual Reliability Metrics Comparison (ZsRE)

4.4 Multilingual Learning to Edit

Do instruction-tuned models generalize Arabic edits cross-lingually The Learning to Edit (LTE) framework (Jiang et al., 2024) was originally proposed to teach English models to incorporate edits through instruction tuning. We extend this framework to support Arabic and multilingual training, evaluating three variants: **LTE-EN**: Trained only on English edits, **LTE-AR**: Trained only on Arabic edits, **LTE-AR-EN**: Jointly trained on Arabic and English edits. We assess both *multilingual performance* (editing and evaluating in the same language) and *cross-lingual performance* (editing in one language, evaluating in another).

Figure 3a compares all methods across reliability, generality, locality, and portability. LTE-AR-EN outperforms all others, showing that joint multilingual training yields the most consistent and robust edit behavior. While LTE-EN performs well in Arabic despite never seeing Arabic edits, adding Arabic fine-tuning further improves generality and reliability. Notably, there is a slight drop in locality for the jointly trained model, reflecting a common trade-off between generalization and specificity.

Figures 3b and 3c further show that LTE fine-tuning substantially improves cross-lingual performance across all metrics, with LTE-AR-EN again achieving the strongest results.

5 Conclusion

We presented the first study of knowledge editing for Arabic, evaluating four editing paradigms: ROME, MEMIT, ICE, and LTE, on the ZsRE and Counterfact benchmarks. Our experiments reveal that parameter-based editing methods, though effective in English, struggle in Arabic and show poor crosslingual transfer. In contrast, instruction-tuned methods, especially our extended multilingual LTE framework, exhibit robust performance both in Arabic and across languages. Our findings highlight key challenges and opportunities in multilingual knowledge editing. First, language-specific morphological and syntactic factors significantly affect the reliability and locality of edits. Second, crosslingual propagation is limited in most existing approaches, emphasizing the need for multilingual training. Finally, instruction tuning emerges as a promising direction for building language-agnostic editing capabilities. We hope this work serves as a foundation for future efforts aimed at scalable and reliable knowledge editing for low-resource and morphologically rich languages like Arabic.

References

Muhammad Asif Ali, Nawal Daftardar, Mutayyba Waheed, Jianbin Qin, and Di Wang. 2025. MQA-KEAL: Multi-hop question answering under knowledge editing for Arabic language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5629–5644, Abu Dhabi, UAE. Association for Computational Linguistics.

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. 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.

Nadir Durrani, Basel Mousi, and Fahim Dalvi. 2025. Editing across languages: A survey of multilingual knowledge editing. In *In The Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (To Appear)*, Suzhou, China. Association for Computational Linguistics.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini, editors. 2024. *Proceedings of the Second Arabic Natural Language Processing Conference*. Association for Computational Linguistics, Bangkok, Thailand.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, Qun Liu, and Wei Wang. 2024. Learning to edit: Aligning LLMs with knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4705, Bangkok, Thailand. Association for Computational Linguistics.

Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, and Kaixiong Zhou. 2024. Cross-lingual multi-hop knowledge editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11995–12015, Miami, Florida, USA. Association for Computational Linguistics.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35 of *NeurIPS*, New Orleans, LA.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. *Preprint*, arXiv:2206.06520.

Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. 2024. Exploring alignment in shared cross-lingual spaces. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6326–6348, Bangkok, Thailand. Association for Computational Linguistics.

Ercong Nie, Bo Shao, Zifeng Ding, Mingyang Wang, Helmut Schmid, and Hinrich Schütze. 2025. Bmike-53: Investigating cross-lingual knowledge editing with in-context learning. *Preprint*, arXiv:2406.17764.

Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham, editors. 2023. *Proceedings of ArabicNLP 2023*. Association for Computational Linguistics, Singapore (Hybrid).

Nianwen Si, Hao Zhang, and Weiqiang Zhang. 2024. Mpn: Leveraging multilingual patch neuron for cross-lingual model editing. *Preprint*, arXiv:2401.03190.

Daniel Tamayo, Aitor Gonzalez-Agirre, Javier Hernando, and Marta Villegas. 2024. Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5831–5847, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-lingual knowledge editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11676–11686, Bangkok, Thailand. Association for Computational Linguistics.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024b. Knowledge editing for large language models: A survey. *Preprint*, arXiv:2310.16218.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024c. 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.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. MLaKE: Multilingual knowledge editing benchmark for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4457–4473, Abu Dhabi, UAE. Association for Computational Linguistics.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. 2023. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *Preprint*, arXiv:2308.09954.

Yuchen Wu, Liang Ding, Li Shen, and Dacheng Tao. 2025. Edit once, update everywhere: A simple framework for cross-lingual knowledge synchronization in llms. *Preprint*, arXiv:2502.14645.

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.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, and 3 others. 2024. A comprehensive study of knowledge editing for large language models. *Preprint*, arXiv:2401.01286.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Multilingual knowledge editing with language-agnostic factual neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

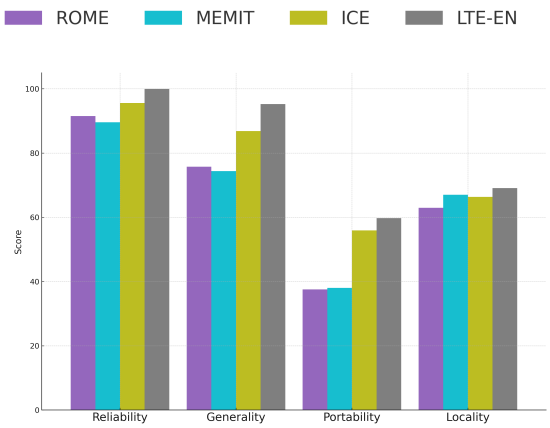


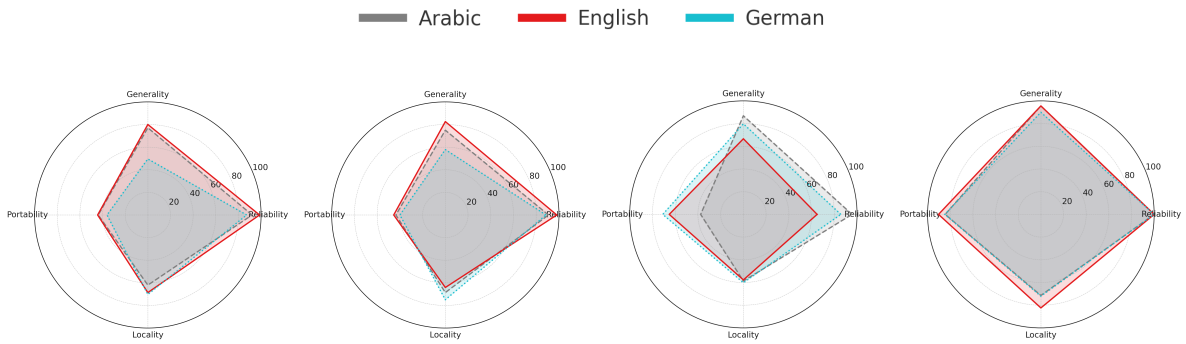
Figure 5: Comparison of ROME, MEMIT, and ICE on LLaMA2-7B-Chat across four metrics: reliability, generality, locality, and Portability on the counterfact dataset

A Additional Results

Arabic Editing The results of Arabic editing performance on the counterfact dataset are shown in figure 5.

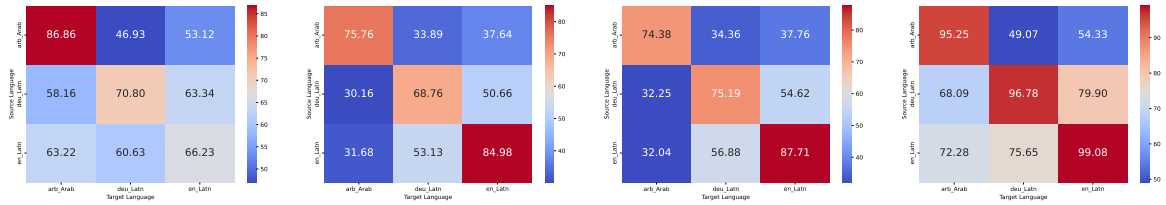
Multilingual Comparison The results of the multilingual comparison on the counterfact dataset are shown in figure 6

Additional Cross-Lingual Results The crosslingual generality metric on the ZsRE are shown in figure 7 and the cross-lingual reliability and generality metric on counterfact are shown in figures 8 & 9



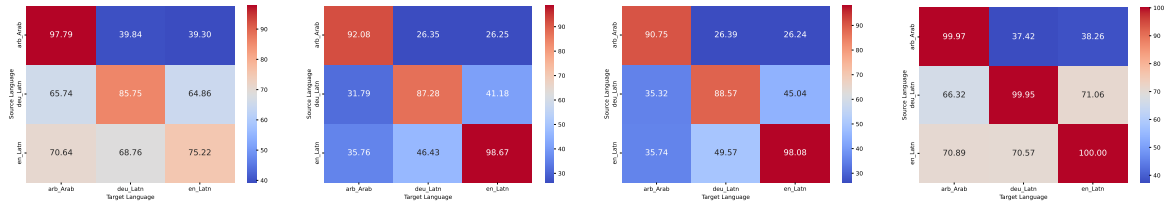
(a) ROME (b) MEMIT (c) ICE (d) LTE

Figure 6: Impact of the editing language on the reliability, generality, portability and locality metrics on counterfact datasets for Llama2-7B-Chat



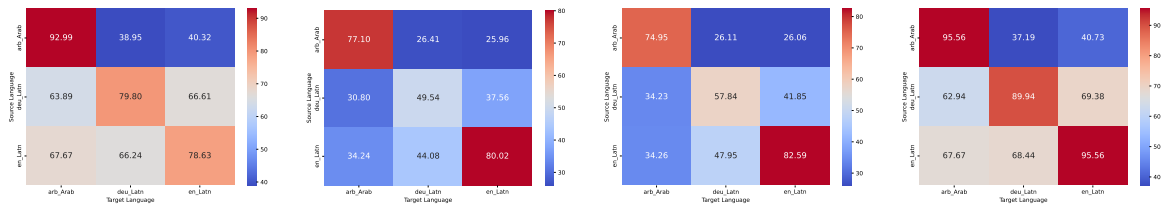
(a) ICE (b) ROME (c) Memit (d) LTE

Figure 7: Cross Lingual Generality Metrics Comparison (ZsRE)



(a) ICE (b) ROME (c) Memit (d) Memit

Figure 8: Cross Lingual Reliability Metrics Comparison (Counterfact)



(a) ICE (b) ROME (c) Memit (d) LTE

Figure 9: Cross Lingual Generality Metrics Comparison (Counterfact)