Editing Across Languages:
A Survey of Multilingual Knowledge Editing

Nadir Durrani[*]   **Basel Mousi**[*]   **Fahim Dalvi**

Qatar Computing Research Institute, HBKU, Doha, Qatar

{ndurrani,bmousi,faimaduddin}@hbku.edu.qa

Abstract

While Knowledge Editing has been extensively studied in monolingual settings, it remains underexplored in multilingual contexts. This survey systematizes recent research on ***Multilingual Knowledge Editing (MKE)***, a growing subdomain of model editing focused on ensuring factual edits generalize reliably across languages. We present a comprehensive taxonomy of MKE methods, covering parameter-based, memory-based, fine-tuning, and hypernetwork approaches. We survey available benchmarks, summarize key findings on method effectiveness and transfer patterns, and identify persistent challenges such as cross-lingual propagation, language anisotropy, and limited evaluation for low-resource and culturally specific languages. We also discuss broader concerns such as stability and scalability of multilingual edits. Our analysis consolidates a rapidly evolving area and lays the groundwork for future progress in editable language-aware LLMs.

1 Introduction

The static nature of pre-trained LLMs presents an important limitation: factual knowledge inevitably becomes outdated or incorrect over time. To address this, the research community has developed knowledge editing (KE) techniques (De Cao et al., 2021; Mitchell et al., 2022b; Meng et al., 2023a) that aim to modify specific factual information within LLMs efficiently, without costly full retraining or broad performance degradation.

While KE methods have made impressive progress in monolingual settings, they overlook a critical real-world dimension: **multilinguality**. Multilingual LLMs are increasingly deployed in cross-lingual applications, and it is insufficient for a knowledge update to take effect only in a single language. Ideally, an edit in one language should propagate across all supported languages, ensuring consistent and equitable access to updated knowledge regardless of linguistic background. This challenge has given rise to the emerging field of **Multilingual Knowledge Editing (MKE)**, which seeks to synchronize factual updates across languages while preserving unrelated knowledge.

In this paper, we present the **first comprehensive survey of methods and benchmarks for multilingual knowledge editing**. We review recent advancements in (i) proposed MKE methodologies, including parameter editing methods, memory-based methods, fine-tuning methods, and hypernetwork-based methods; (ii) benchmark datasets designed to evaluate multilingual knowledge editing effectiveness and cross-lingual generalization; (iii) empirical findings on method effectiveness and cross-lingual generalization; and (iv) key open challenges and future directions toward building LLMs that are reliably editable across languages. Our goal is to provide a structured foundation for future research and help unify this fast-growing but currently scattered domain.

While several surveys have examined knowledge editing in LLMs, they focus primarily on monolingual settings or broad architectural taxonomies. Mazzia et al. (2024) provide a general overview of neural model editing across modalities. Wang et al. (2024c) emphasize technical classifications and optimization formulations, while Zhang et al. (2024a) focus on monolingual editing and interpretability. None of these surveys consider the specific challenges of multilingual propagation or evaluation.

By contrast, our survey focuses explicitly on the multilingual dimension. We contribute (i) a multilingual reinterpretation of existing method families, (ii) a synthesis of empirical trends observed across MKE studies, and (iii) targeted visualizations comparing methods, languages, and model scales. Our work complements and extends prior surveys by highlighting multilingual knowledge editing as a critical but underexplored area.

---

[*]The authors contributed equally

2 Multilingual Knowledge Editing

Large Language Models (LLMs) are parameterized functions $f(x; \theta)$, where θ denotes the model weights. While LLMs acquire extensive factual knowledge during pre-training, this knowledge remains fixed and may become outdated or incorrect.

Knowledge Editing (KE) modifies a model's behavior to reflect updated facts, either by changing its parameters θ or by influencing outputs through prompts or external memory. The objective is to apply edits locally, affecting only the intended fact while preserving unrelated knowledge.

This survey focuses on **Multilingual Knowledge Editing (MKE)**, where edits introduced in one language (e.g., English) should generalize to semantically equivalent queries in other languages, without disrupting unrelated outputs.

Let $\mathcal{L} = \{\ell_1, \ell_2, \ldots\}$ denote the set of languages, and represent a fact as a triple (s, r, o_{new}). Given queries Q_{ℓ_s} in a source language ℓ_s, and corresponding queries Q_{ℓ_t} in target languages $\ell_t \in \mathcal{L}$, the goal is to find updated parameters θ'[1] such that:

$$f(q; \theta') \approx o_{\text{new}} \quad \text{for } q \in Q_{\ell_s} \cup Q_{\ell_t},$$
$$f(q; \theta') \approx f(q; \theta) \quad \text{for } q \notin Q_{\ell_s} \cup Q_{\ell_t}.$$

A key challenge in MKE is *language anisotropy*, where multilingual models encode different languages in misaligned subspaces, making edits difficult to propagate reliably. In this survey, we define the multilingual knowledge editing problem and highlight its core challenges. The following sections cover proposed MKE methods, benchmark datasets, empirical findings across models and languages, and open research directions.

3 Methods in Multilingual KE

Work in multilingual knowledge editing (MKE) builds on monolingual techniques and introduces new methods to support cross-lingual consistency. These approaches can be broadly grouped into four methodological families: (i) **Parameter Editing**, which modifies model weights directly; (ii) **Memory-based** methods, which use in-context learning or external memory without updating parameters; (iii) **Fine-tuning**, including full and instruction-tuned updates; and (iv) **Hypernetwork-based**, which generate edits via auxiliary networks. This taxonomy supports a structured comparison of design choices and trade-offs in MKE.

---

[1]Non-parametric methods apply edits at inference time through prompts or external memory without modifying θ.

3.1 Parameter Editing Methods

Parameter editing methods modify the internal weights of LLMs to update factual knowledge. They can be broadly divided into two categories: general methods originally developed for monolingual settings and later applied to multilingual models, and specialized methods designed to support cross-lingual propagation of edits.

3.1.1 General KE Methods Adapted for MKE

ROME (Meng et al., 2023a) introduced the locate-then-edit paradigm, identifying key activation pathways and applying a low-rank update to model weights. **MEMIT** (Meng et al., 2023b) extended ROME to support batch editing of multiple facts using rank-one updates $\{\Delta\theta_k\}_{k=1}^K$ distributed across layers. While it preserves locality and improves efficiency, its effectiveness across languages is limited by language-specific neuron activations. Similarly, **KnowledgeNeuron** (Dai et al., 2022) and **PMET** (Li et al., 2024) followed the locate-and-edit paradigm by optimizing $\Delta\theta$ to encode factual edits. Wang et al. (2024a); Zhang et al. (2025) adapted **ROME**, **MEMIT**, and **PMET** to multilingual models. While these adaptations preserve edit locality and maintain computational efficiency, they suffer from poor cross-lingual propagation due to language-specific activation patterns and representational anisotropy.

3.1.2 Specific Methods Designed for MKE

MEMAT (Tamayo et al., 2024) incorporated attention head analysis into the locate-then-edit paradigm to improve cross-lingual generalization. It ranks heads based on their contribution to predictions in target languages and applies targeted corrections. **MPN** (Si et al., 2024) introduced patch neurons, key-value units trained to jointly encode facts across languages. These patches update knowledge without modifying original weights, promoting multilingual consistency. **LU-LAFNs** (Zhang et al., 2025) takes a different approach by identifying shared high-attribution neurons across languages and caching their values to apply edits dynamically at inference, enabling non-destructive, language-independent interventions. Finally, **MEMLA** (Xie et al., 2024) used integrated gradients to locate language-relevant neurons and applies masked LoRA updates to targeted subsets, balancing cross-lingual generalization and edit locality.

3.1.3 Summary

Strengths: Parameter editing methods offer localized updates that reduce interference with unrelated knowledge and are efficient for factual edits without full retraining. They are also compatible with standard transformer architectures, making them easy to integrate into multilingual LLM pipelines.

Weaknesses: These methods rely on shared internal representations across languages, which are often weak or inconsistent. As a result, they do not guarantee reliable cross-lingual propagation; edits made in one language may fail to transfer to others. Their performance is also sensitive to tokenization and neuron alignment, particularly in morphologically rich or low-resource languages. While MKE-specific approaches (e.g LU-LAFNs) address these issues through attention or neuron-level interventions, key theoretical challenges remain.

3.2 Memory-based Methods

Memory-based methods perform knowledge editing without modifying model weights θ, using either in-context interventions or external memory to override factual outputs. These approaches offer greater flexibility and reduce the risk of unintended interference. We group them into two groups: **In-context Editing** and **Memory-Retriever** methods.

3.2.1 In-Context Editing (ICE & IKE)

In-context editing (Zheng et al., 2023) modifies model behavior by injecting factual corrections into the input prompt, without changing model weights. Zero-shot ICE uses single factual statements, while IKE extends this with few-shot demonstrations. These methods can generalize across languages if equivalent demonstrations are available in each language. Nie et al. (2025) for example applied in-context editing in 53 languages.

However, these methods are constrained by context length: edits must be carried in the prompt, which reduces space for task inputs and outputs. This issue is especially severe in languages with high tokenizer fertility, where facts occupy more tokens, and in sequential editing scenarios, where multiple edits quickly accumulate.

3.2.2 Memory-Retriever Editing

Memory-Retriever methods attach an external memory to the LLM and retrieve language-specific facts $\{m_i\}$ at inference time, producing $f(q, \{m_i\}; \theta) \approx o_{\text{new}}$ without modifying weights.

This design supports multilingual edits by isolating updates outside the model, but effectiveness depends on retriever quality. Variants differ in how memory is used: **SERAC** (Mitchell et al., 2022b) overrides predictions with matched entries. **ReMaKE** (Wang et al., 2024d) uses retrieved memories to guide lightweight updates. **MQA-KEAL** (Ali et al., 2025) performs multi-hop reasoning over memory chains. Cross-lingual retrievers like **Mello-CL** (Khandelwal et al., 2024) boost recall in multilingual and low-resource contexts.[2]

3.2.3 Summary

Strengths: Memory-based methods avoid parameter updates, minimizing interference with unrelated cross-lingual knowledge. Their edits are applied externally at inference time, enabling high flexibility for dynamic and reversible updates. Because corrections are decoupled from internal weights, these approaches can be applied independently to any language, making them naturally suitable for multilingual deployments.

Weaknesses: Their effectiveness depends on language-specific components: prompt understanding and instruction following for in-context methods, or retriever accuracy and memory quality for retrieval-based approaches. Neither guarantees cross-lingual propagation; corrections must be explicitly provided or retrieved in each target language. In addition, memory-retriever methods incur storage and latency costs as edits accumulate, while in-context editing requires constructing effective prompts across languages.

3.3 Fine-Tuning Methods

Fine-tuning updates large parameter blocks, such as full weights or adapters, to encode factual edits. Unlike parameter editing, which operates on localized parameter subspaces, it offers stronger reliability but higher risk of side effects in multilingual settings. These methods fall into two types: *baseline fine-tuning*, which trains directly on factual updates, and *instruction-tuned fine-tuning*, which treats editing as a task via structured instructions.

---

[2]Other approaches such as **GRACE** (Hartvigsen et al., 2023) and **WISE** (Wang et al., 2024b) also support handling multiple edits while maintaining locality. We did not include them in our main taxonomy since they were developed and evaluated only in English and do not address cross-lingual generalization, but they represent promising directions for future MKE research.

3.3.1 Baseline Fine-Tuning Methods

Baseline fine-tuning methods apply supervised learning to update model parameters with edited knowledge. **Full Fine-Tuning (FT)** updates all model parameters θ by minimizing loss over edited facts, making it a simple and commonly used baseline. **LoRA-FT** (Hu et al., 2021) introduces low-rank adaptation matrices into transformer blocks and updates only these, offering a parameter-efficient alternative. **FT-L** constrains updates to selected layers using gradient masking (Zhu et al., 2020), while **FT-M** fine-tunes a specific MLP sublayer. These methods, originally developed for monolingual editing, are applied as baselines (e.g. in Wu et al. (2025)) for multilingual knowledge editing. While they offer strong edit reliability in the source language, generalization to target languages remains purely implicit, with no mechanism to enforce cross-lingual propagation.

3.3.2 Instruction-Tuned Knowledge Editing

Instruction-tuned methods frame editing as a task, training models to apply factual updates through structured edit instructions rather than direct parameter overwriting. They are fine-tuned on supervised pairs of edit requests and desired outputs, enabling the model to internalize editing behavior.

Learning to Edit (LTE) (Jiang et al., 2024) generates synthetic edits and fine-tunes the model to process an edit descriptor E and query q to output:

$$f(q, E; \theta') \approx o_{\text{new}} \quad \forall q \in Q_{\ell_s}.$$

X-KDE (Wu et al., 2025) extended LTE to multilingual settings via a two-stage framework: Cross-lingual Edition Instruction Tuning (XE-IT) trains on parallel data to transfer edits across languages, and Target-Language Preference Optimization (TL-PO) aligns outputs for consistency.

$$f(q, E; \theta') \approx o_{\text{new}} \quad \forall q \in Q_{\ell_s} \cup Q_{\ell_t}, \ell_t \in \mathcal{L}.$$

Instruction-tuned approaches are the first systematic attempt to model editing as a generalizable capability. LTE is effective in monolingual contexts, while X-KDE introduced a supervised framework for cross-lingual propagation and alignment.

3.3.3 Summary

Strengths: Fine-tuning methods are widely used for knowledge editing due to their simplicity and strong factual reliability. They are compatible with pre-trained transformer architectures and require minimal structural changes. LoRA-based fine-tuning improves efficiency by updating a small parameter subset. Instruction-tuned approaches, such as LTE and X-KDE, treat editing as a learnable task, offering more control over behavior and reducing unintended side effects. Notably, X-KDE is the first supervised fine-tuning framework specifically designed to promote cross-lingual consistency.

Weaknesses: Fine-tuning provides no guarantees for cross-lingual propagation; generalization to target languages depends on implicit internal alignment. Large parameter updates also risk altering unrelated knowledge. While instruction-tuned methods mitigate this by explicitly modeling the editing process, they rely on large, curated datasets and careful instruction design.

3.4 Hypernetwork-Based Methods

Hypernetwork-based methods edit knowledge by training auxiliary networks to generate updates. Instead of manual weight changes, hypernetwork learns to produce edits from query-output pairs. **MEND** (Mitchell et al., 2022a) introduced this by training a hypernetwork h_ϕ to produce updates $\Delta\theta = h_\phi(q, o_{\text{new}})$ such that $f(q; \theta + \Delta\theta) \approx o_{\text{new}}$, while minimizing interference. **KnowledgeEditor** (De Cao et al., 2021) added regularization to stabilize edits and improve generalization. **LiME** (Xu et al., 2023) adapted this paradigm for multilingual editing by conditioning the hypernetwork on source language ℓ_s, generating edits in language-specific subspaces: $f(q; \theta + h_\phi(q, o_{\text{new}}, \ell_s)) \approx o_{\text{new}}$. By modeling language anisotropy explicitly, LiME improved cross-lingual stability.

3.4.1 Summary

Strengths: Hypernetwork-based methods enable meta-learning of edits without manual parameter selection, offering fast, localized updates with minimal disruption. Language-conditioned variants such as LiME introduce explicit mechanisms to improve edit stability across languages.

Weaknesses: The effectiveness of hypernetwork-based methods depends heavily on the quality and coverage of training data. While **MEND** and **KnowledgeEditor** were designed for monolingual models and used as multilingual baselines, Beniwal et al. (2024) showed their limited cross-lingual transfer. **LiME** mitigates this by conditioning on the source language, but its reliance on language-specific subspaces may limit generalization across distant language pairs.

Method	Edit Rel	X-ling Prop	Locality	Param. Mod	Dyn. Flex
Parameter Editing Methods					
ROME	✓	✗	✓	✓	✗
MEMIT	✓	✗	✓	✓	✗
MEMAT	✓	~	✓	✓	✗
MPN	✓	~	✓	✓	✗
LU-LAFNs	✓	~	✓	✗	✓
MEMLA	✓	~	✓	✓	✗
Memory-Based Methods					
IKE / ICE	~	~	✓	✗	✓
SERAC	✓	✗	✓	✗	✓
ReMaKE	✓	~	✓	✗	✓
MQA-KEAL	✓	✗	✓	✗	✓
MELLO-CL	~	~	✓	✗	✓
Fine-Tuning Methods					
FT	✓	✗	✗	✓	✗
LoRA-FT	✓	✗	~	✓	✗
FT-L / FT-M	✓	✗	~	✓	✗
LTE	✓	✗	~	✓	✗
X-KDE	✓	~	~	✓	✗
Hypernetwork-Based Methods					
MEND	✓	✗	✓	✓	✗
KnowledgeEditor	✓	✗	✓	✓	✗
LiME	✓	~	✓	✓	✗

Table 1: Comparison of multilingual knowledge editing methods across core evaluation criteria. **Edit Rel**: edit reliability in source language; **X-ling Prop**: cross-lingual propagation; **Locality**: preservation of unrelated knowledge; **Param. Mod**: modifies model weights; **Dyn. Flex**: supports inference-time edit control. ✓ = strong; ∼ = partial; ✗ = absent.

3.5 Summary of Methods

Table 1 summarizes the trade-offs across multilingual knowledge editing methods. Parameter editors offer strong locality and reliability but struggle with propagation. Memory-based methods are flexible and reversible, yet rely on language-specific retrieval or prompting. Fine-tuning provides reliable edits but lacks adaptability. Instruction-tuned approaches such as X-KDE improve generalization with explicit supervision. Hypernetwork methods like LiME minimize interference but remain underexplored in multilingual contexts. No method meets all requirements, motivating further research into hybrid strategies.

4 Evaluation for Multilingual KE

Despite recent progress in multilingual knowledge editing (MKE), the field still lacks a unified evaluation framework. Most studies build on protocols developed in the monolingual KE literature, typically using four core criteria:

Reliability measures whether the edit is successful in the source language, typically quantified as the percentage of source-language queries that yield the updated fact.

Generality evaluates robustness to linguistic variation, assessing model accuracy on paraphrased or rephrased queries in the source language.

Benchmark	Langs	Human?	Notes
MLaKE (Wei et al., 2025)	5	✓	ChatGPT-generated, human-verified; no paraphrase/locality samples
MzSRE (Wang et al., 2024d)	12	✗	Translated via Google; used for all four metrics
biZsRE (Wang et al., 2024a)	2	~	GPT-translated; test set verified by native speakers
BMIKE-53 (Nie et al., 2025)	53	~	GPT4o-translated; BLEU + manual inspection
MQuAKE (Khandelwal et al., 2024)	7	✓	Translations verified via BLEU and human checks
MQuAKE-AR (Ali et al., 2025)	1	✓	LLM-based scoring with manual refinement
MKEB (Xie et al., 2024)	12	✗	SPARQL + Baidu; no human revision
DocTer (Wu et al., 2023)	2	?	Generated from Counterfact using ChatGPT
Counterfact (CA) (Tamayo et al., 2024)	2	✓	Aina-translated; SimAlign + human filtering
CKnowEdit (Fang et al., 2024)	1	✓	Chinese cultural and linguistic error correction

Table 2: Multilingual knowledge editing benchmarks. Langs. = number of languages. Human? = human annotation or translation verification.

Locality examines how unrelated knowledge is preserved, usually by comparing model predictions on non-target queries before and after the edit.

Portability captures the extent to which an edit generalizes beyond its original setting, including transfer to downstream tasks, multi-hop reasoning chains, or different model variants.

In the multilingual setting, these dimensions are extended to account for *cross-lingual propagation*, whether an edit made in one language transfers accurately to semantically equivalent queries in other languages. This multilingual extension has become a defining challenge in MKE, and features prominently in recent work (Wei et al., 2025; Zhang et al., 2025; Wu et al., 2025; Xie et al., 2024).[3]

Some recent work (Ali et al., 2025; Khandelwal et al., 2024) further expands evaluation by introducing *multi-hop reasoning* over edited knowledge, using variants of the multi-hop accuracy metric introduced by Zhong et al. (2023). These setups test whether edits are faithfully integrated into the model's broader knowledge base and whether they support multilingual inference chains.

To support evaluation, a growing number of multilingual benchmarks have emerged. These datasets vary in language coverage, task complexity (e.g., single-hop vs. multi-hop), and translation quality, reflecting the diversity of MKE settings. Table 2 summarizes key characteristics. Additional details on construction and evaluation are in Appendix A.

---

[3] Terminology varies across studies. For example, Xie et al. (2024) refer to reliability as the efficacy score, generality as the paraphrase score, and locality as the neighborhood score.
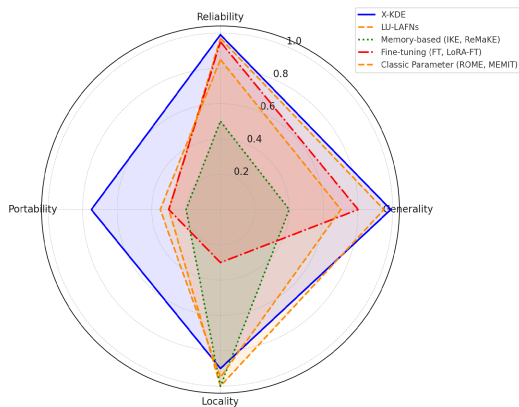
Figure 1: Comparison of multilingual knowledge editing methods across four evaluation criteria: reliability, generality, locality, and portability. Methods are grouped and color-coded by method family. Scores represent an approximate synthesis of reported results from recent studies (Zhang et al., 2025; Wu et al., 2025)

5 Findings

We now highlight recurring trends and differences across method families, focusing on how these methods perform under the evaluation dimensions discussed in the previous section.

5.1 Comparative Evaluation

Q: Which methods best balance reliability, generality locality and portability in MKE?

Experimental comparisons of multilingual knowledge editing (MKE) methods reveal patterns in how different approach families trade off reliability, generality, locality, and portability.

Parameter Editing Parameter-based methods such as ROME and MEMIT perform strongly in monolingual settings but struggle in multilingual contexts, often causing interference and degraded cross-lingual transfer (Wang et al., 2024d; Wu et al., 2025; Zhang et al., 2025; Wei et al., 2025; Xie et al., 2024). LU-LAFNs achieves perfect locality and strong reliability by identifying and modifying shared factual neurons across languages. A key reason for its superior locality is how updates are applied: after optimization, the updates are stored externally and only injected at inference time if a matching subject is detected in the user prompt. This minimizes the risk of unintended changes to unrelated knowledge. However, like other parameter editing methods, LU-LAFNs shows weaker cross-lingual generalization and portability.

Memory-based Methods Memory-based methods like IKE and ReMaKE avoid parameter up-

dates and instead rely on retrieval or prompting. ReMaKE achieves excellent locality but suffers from low reliability in many multilingual scenarios, with performance often dropping below 50%. In contrast, IKE demonstrates relatively poor locality, as it heavily depends on the quality and selection of few-shot exemplars (Wang et al., 2024a,d; Wu et al., 2025). Additionally, the applicability of ReMaKE across model families remains limited, its performance degrades significantly when applied to models such as BLOOM (Zhang et al., 2025), indicating strong dependence on the underlying model's capabilities. Despite their limitations, prompting-based approaches remain surprisingly competitive: Nie et al. (2025) demonstrated that simple in-context editing achieves reasonable cross-lingual edit success across 53 languages, even without model weight modifications.

Fine-tuning Fine-tuning strategies (e.g., FT-L, FT-M, LoRA-FT) deliver high reliability for monolingual edits but perform worst in locality, often degrading unrelated knowledge (Xu et al., 2023; Wu et al., 2025). They also exhibit limited ability to transfer edits across languages, making them the least favorable option for MKE in general. An important exception is X-KDE (Wu et al., 2025), which combines fine-tuning with cross-lingual instruction tuning and preference optimization to achieve the best reported overall balance of reliability, generality, and locality across languages. Among fine-tuning methods without explicit adaptation for MKE, LTE stands out for its strong cross-lingual generalization despite being trained only on monolingual synthetic edits (Wu et al., 2025).

Hypernetwork-based Methods No large-scale multilingual benchmarking of hypernetwork-based methods exists. Early results in monolingual and small-scale multilingual settings suggest potential for efficient edits and strong generalization, though systematic evidence remains limited.

Summary Current evidence consistently identifies **X-KDE as offering the best overall balance of reliability, generality, and locality across languages. LU-LAFNs stands out within parameter editing methods** for its unmatched locality through targeted neuron editing and conditional application of updates, although it exhibits weaker cross-lingual portability. Classic parameter editors and fine-tuning approaches perform substantially worse in cross-lingual settings, and memory-based meth-
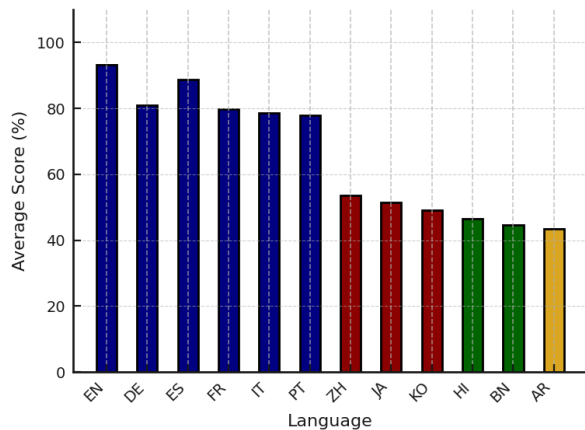
Figure 2: Average Reliability and Generality for MEMLA (Xie et al., 2024) across twelve languages, illustrating a typical directionality pattern: edits in high-resource or related languages transfer better than those in low-resource or distant ones.

ods, while safe in terms of locality, offer limited reliability and portability. Hypernetwork-based methods remain promising but underexplored in the context of multilingual knowledge editing. Figure 1 provides a visual summary of the relative performance of these method families across the four evaluation criteria.

5.2 Cross-Lingual Transferability

Q: What factors influence the cross-lingual propagation of edits in MKE, and how well do current methods address this challenge?

Experimental studies on MKE consistently show that cross-lingual transfer remains a major challenge. Analysis of recent benchmarks highlights several key factors influencing the effectiveness of cross-lingual propagation of edits.

Language Relatedness   The strongest predictor of successful cross-lingual propagation is the degree of linguistic similarity between the source and target languages. Edits propagate better within the same language family (e.g., English to Spanish or German (see Figure 2)) than between distant language families (e.g., English to Chinese or Arabic) (Mousi et al., 2025; Zhang et al., 2025; Wei et al., 2025). Some studies further suggest that shared writing systems or similar orthographic structures (e.g., Latin-based scripts) may offer marginal additional benefits to transferability (Nie et al., 2025; Xu et al., 2023). Beniwal et al. (2024) also showed that script differences (e.g., Devanagari vs. Tamil) reduce edit transferability, highlighting script divergence as a secondary but relevant

limiting factor. Recent work further underscores the role of subject token similarity in transfer success: Tamayo et al. (2024) showed that edits are more likely to propagate when the embedding representations of subject entities are closely aligned across languages.

Language and Directionality of Transfer   The language in which the edit is applied has a notable impact on cross-lingual propagation. Edits performed in high-resource languages such as English generally transfer more effectively to other languages than edits made in lower-resource or non-Latin languages such as Arabic (Mousi et al., 2025) or Chinese (Wu et al., 2025; Zhang et al., 2025; Xie et al., 2024; Nie et al., 2025). This asymmetry likely reflects the dominant role of English and other high-resource languages during pretraining and fine-tuning of multilingual language models.

Challenges in Multi-hop Reasoning   Transferability further degrades in multi-hop reasoning settings, where complex fact chains must be updated across multiple languages (Wei et al., 2025; Xie et al., 2024). Existing methods show significant performance drops in multi-hop multilingual scenarios compared to single-hop edits.

Overall Limitations   Despite recent progress, no method achieves fully consistent and reliable cross-lingual propagation across all language pairs. Multiple studies report large variability in performance, especially for distant and low-resource languages (Nie et al., 2025; Wei et al., 2025; Xie et al., 2024). Cross-lingual transferability thus remains an open challenge for the field.

5.3 Impact of Model Size and Capability

Q: How do model size affect the performance of multilingual knowledge editing?

Larger and more capable models consistently yield higher editing quality across reliability, generality, locality, and, to a lesser extent, portability. LLaMA-3.1-8B outperforms Qwen2-7B and Bloomz-7B, while Qwen2.5-7B-Instruct surpasses Chinese-LLaMA-7B in cross-lingual benchmarks (Wu et al., 2025; Zhang et al., 2025).

An interesting secondary finding is that stronger models also stabilize weaker editing methods. For example, fine-tuning performs poorly on weaker models like Bloomz but becomes stable, though still suboptimal on LLaMA-3 and Qwen2 (Zhang
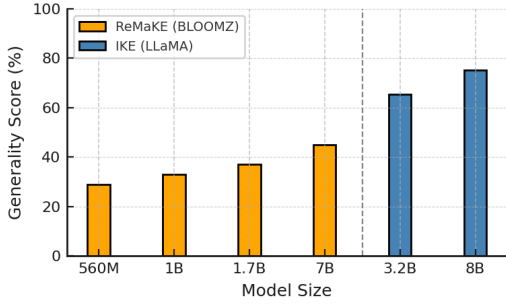
Figure 3: Generality scores for BLOOMZ and LLaMA models across increasing model sizes. ReMaKE (Wang et al., 2024d) (BLOOMZ) results are extracted for mzsRE generality, averaged across 10 languages. IKE (Nie et al., 2025) (LLaMA) results are extracted for mzsRE generality under zero-shot settings, averaged across 53 languages. Results show that generality consistently improves with model scale for both models.

Method Family	Sequential	Batch	Notes
Parameter Editing	✗ (ROME) ✓ (LU-LAFNs) ∼ (MEMLA)	✗ (ROME) ✓ (MEMIT, MEMLA, LU-LAFNs)	MEMIT supports batch; LU-LAFNs allow partial sequential edits.
Memory-Based	✓	✓	Edits stored externally; supports dynamic updates and reversibility.
Fine-Tuning	✗ (FT, LoRA-FT) ✓ (X-KDE, LTE)	✗ (FT, LoRA-FT) ∼ (X-KDE, LTE)	High cost per update; X-KDE enables reusable batch edits.
Hypernetwork-Based	✓	✓	Meta-learned edits; LiME supports low-latency, language-conditioned updates.

Table 3: Compatibility of MKE method families with sequential and batch editing. ✓ = supported; ∼ = partial; ✗ = not supported.

et al., 2025). Parameter editors like M-MEMIT show similar improvements on stronger models.

Instruction tuning further enhances robustness. Models such as Qwen2.5-7B-Instruct and LLaMA-chat exhibit slightly better generality and portability as shown in Wu et al. (2025). This suggests instruction-tuned models may possess a more disentangled and structured internal knowledge space, which aids editing consistency.

Finally, stronger models reduce the performance gap across methods. On Bloomz-7B, the performance difference between LU-LAFNs and M-MEMIT exceeds 25 points, whereas on LLaMA-3.1-8B the gap shrinks to under 10 points (Zhang et al., 2025). However, the hardest remaining challenge is portability. Even strong models struggle to propagate edits across English and Chinese, with only X-KDE and LU-LAFNs maintaining acceptable performance in this setting (Wu et al., 2025).

6 Discussion

Multilingual knowledge editing has made progress, but major limitations remain. In this section, we identify persistent challenges and outline opportunities for future work. We organize these into thematic groups, reflecting both empirical trends and conceptual gaps in the literature.

6.1 Open Challenges

Language Anisotropy and Representation Misalignment Multilingual LLMs encode languages into overlapping yet partially distinct internal subspaces (Mousi et al., 2024), creating structural anisotropy that can hinder cross-lingual trans-

fer. A key open problem for MKE is to bridge these spaces by using shared regions and handling language-specific pockets.

Lack of Evaluation Coverage for Low-Resource Languages Most MKE studies evaluate only high-resource languages (e.g., English, Chinese). Systematic evaluation on low-resource, morphologically complex, and under-represented languages (e.g., Swahili, Urdu) remains largely absent. In addition, current MKE benchmarks largely assume that factual knowledge is language-invariant. This overlooks culturally specific knowledge that may not have direct equivalents across languages, which is especially important for low-resource languages where such references are more prominent. Developing evaluation resources that account for cultural grounding remains an open challenge.

Trade-off Between Locality and Propagation Methods that enforce strong locality (e.g., LU-LAFNs) tend to sacrifice propagation, while those that encourage transfer (e.g., fine-tuning, prompting) risk damaging unrelated knowledge. Designing architectures or training objectives that mitigate this tension remains an open research direction.

Stability and Model Collapse Another critical challenge, underexplored in multilingual settings, is model collapse, where even a few edits can substantially degrade an LLM's general performance (Green et al., 2025). Monolingual studies, (Yang et al., 2024; Gupta et al., 2024) show that edits may trigger instability, and multilingual scenarios, with misaligned representational spaces, exacerbate this risk. Evaluating MKE methods should therefore consider both the success of edits and the preservation of model capabilities, particularly under sequential or batched updates where collapse risks accumulate.

Sequential and Batch Editing Realistic deployment requires models to support not only isolated

edits but also sequential updates and batched revisions. Most current MKE methods are not evaluated in these conditions. Memory-based and hypernetwork approaches (e.g., ReMaKE, LiME) naturally support sequential updates via externalized edits, while parameter editors such as MEMIT suit batch updates. Instruction-tuned methods (e.g., X-KDE) enable reusable edit behavior in batches but incur higher training costs. Table 3 summarizes relative compatibility across method families.

Benchmark Fragmentation and Lack of Standardization   The field still lacks unified evaluation protocols and benchmarks. While MLaKE and MZSRE are important steps forward, inconsistencies in construction, metrics, and coverage persist. Many rely on machine translation outputs, introducing noise. Low-resource and morphologically rich languages remain underrepresented.

Domain-specific Editing   Current MKE research is almost entirely evaluated on general-purpose factual knowledge. The impact of editing in specialized domains such as biomedical, legal, or culturally grounded knowledge remains unexplored. Developing benchmarks and methods for domain-specific MKE would broaden applicability and expose new challenges in precision and transfer.

Scalability and Real-World Applicability   Multilingual knowledge editing methods face key barriers to deployment. Memory-based approaches incur high storage or retrieval costs when scaling across languages. Fine-tuning is computationally expensive, especially when edits must generalize to variants. Parameter editors, though efficient, often lack robustness across diverse language families.

6.2 Opportunities

Development of Robust Multilingual Benchmarks   Future benchmark design must move beyond machine translation and include high-quality, human-curated datasets covering low-resource languages and morphologically rich targets. Multihop reasoning tasks (e.g., MQA-KEAL) should be expanded to test compositional edit capabilities.

Language-Conditioned or Language-Aware Editing   Early work such as LiME shows that explicitly conditioning edits on language identifiers can help mitigate representation anisotropy. Future models could leverage joint multilingual objectives or hierarchical representations to improve robustness across language pairs.

Instruction-Tuned and Task-Oriented Editing
Instruction-tuned paradigms like X-KDE offer an attractive framework to model knowledge editing as a task. Continued exploration of instruction formats, task decomposition, and preference optimization could improve multilingual generalization.

Multilingual In-Context Learning and Adapter Fusion   Combining prompting with adapter-based updates could enable lightweight multilingual editing with minimal model disruption.

Multi-source Editing   Most current MKE methods assume monolingual edits, with cross-lingual propagation only evaluated after the fact. A few approaches (e.g., LU-LAFNs, LiME) explicitly support multi-source editing by learning language-agnostic representations, allowing the same fact to be updated jointly across languages. This capability remains underexplored but represents an important opportunity for ensuring consistency and minimizing interference in multilingual models.

Cross-Model Editing and Transfer   A largely unexplored avenue is whether knowledge edits can transfer across model families. Building model-agnostic edit representations could open new possibilities for scalable MKE.

6.3 Connections to Adjacent Research Areas

Hallucination mitigation   Knowledge editing has been explored as a way to correct hallucinations by directly updating model representations, with benchmarks such as HalluBench (Huang et al., 2025) and TruthX (Zhang et al., 2024b). However, evaluation remains English-only, and studies on multilingual hallucinations focus on measuring or reducing rates without using editing. Thus, the intersection of hallucination mitigation and MKE represents an exciting frontier for future research.

Mechanistic interpretability and cross-lingual transfer   Many editing methods, particularly parameter-based ones, build on mechanistic interpretability, assuming factual knowledge can be localized in neural circuits. Yet little is known about whether such circuits align across languages or how misalignment affects propagation. Exploring this link could ground MKE theoretically and explain why some edits generalize cross-lingually while others remain language-specific. Together, these connections show that MKE both informs and benefits from adjacent areas, though most current work remains monolingual.

Acknowledgments

Limitations

While this survey provides a comprehensive overview of recent advances in multilingual knowledge editing (MKE), our work has several limitations:

- **Cross-paper comparability:** Our synthesis relies on reported results from prior studies. Due to variations in evaluation protocols, dataset versions, and metric definitions, exact comparisons across methods are not always possible. We therefore emphasize relative trends rather than absolute rankings.

- **No new experimental validation:** We do not contribute new experimental results. Our findings on limitations, especially regarding low-resource and under-represented languages, are based on the coverage and reporting of the original papers.

- **Selection completeness:** While we aimed to cover the most influential and representative papers across major method families, the rapidly evolving nature of multilingual knowledge editing means that relevant recent or niche approaches may have been unintentionally omitted.

Ethical Consideration

This work is a survey of existing research on multilingual knowledge editing and does not involve any new data collection, user studies, or model deployment. We do not foresee any direct ethical concerns. All referenced works have been appropriately cited, and any datasets or models discussed were originally released under appropriate terms by their respective authors. AI tools were used to rephrase and improve exposition of sections of the paper.

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.

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.

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.

Jizhan Fang, Tianhe Lu, Yunzhi Yao, Ziyan Jiang, Xin Xu, Ningyu Zhang, and Huajun Chen. 2024. Cknowedit: A new chinese knowledge editing dataset for linguistics, facts, and logic error correction in llms. *Preprint*, arXiv:2409.05806. ArXiv preprint arXiv:2409.05806.

Tommaso Green, Félix Gaschi, Fabian David Schmidt, Simone Paolo Ponzetto, and Goran Glavaš. 2025. BabelEdits: A benchmark and a modular approach for robust cross-lingual knowledge editing of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8342–8369, Vienna, Austria. Association for Computational Linguistics.

Akshat Gupta, Sidharth Baskaran, and Gopala Anumanchipalli. 2024. Rebuilding ROME : Resolving model collapse during sequential model editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21738–21744, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2025. Can knowledge editing really correct hallucinations? In *The Thirteenth International Conference on Learning Representations*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

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.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024. Pmet: Precise model editing in a transformer. *Preprint*, arXiv:2308.08742.

Vittorio Mazzia, Alessandro Pedrani, Andrea Caciolai, Kay Rottmann, and Davide Bernardi. 2024. A survey on knowledge editing of neural networks. *Preprint*, arXiv:2310.19704.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023a. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. *Preprint*, arXiv:2110.11309.

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

Basel Mousi, Nadir Durrani, and Fahim Dalvi. 2025. An exploration of knowledge editing for arabic. *Preprint*, arXiv:2507.09629.

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.

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.

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.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024b. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024c. Knowledge editing for large language models: A survey. *ACM Comput. Surv.*, 57(3).

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

Jiakuan Xie, Pengfei Cao, Yuheng Chen, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Memla: Enhancing multilingual knowledge editing with neuron-masked low-rank adaptation. *Preprint*, arXiv:2406.11566.

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.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024. The butterfly effect of model editing: Few edits can trigger large language models collapse. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5419–5437, Bangkok, Thailand. 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. 2024a. A comprehensive study of knowledge editing for large language models. *Preprint*, arXiv:2401.01286.

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

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.

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

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *Preprint*, arXiv:2012.00363.

A Details on MKE Benchmarks

A range of multilingual benchmarks have been proposed to evaluate MKE methods. Below, we summarize each dataset's design, construction process, and evaluation focus.

MLaKE (Wei et al., 2025) is a benchmark comprising 4,072 multihop and 5,360 single-hop questions across five languages: English, Chinese, Japanese, French, and German. The dataset was constructed by aligning fact chains from Wikipedia, generating raw data via ChatGPT, and finalizing question-answer pairs through human verification. While MLaKE supports both single-hop and multihop evaluation, it lacks paraphrase queries and locality test samples.

MzSRE (Wang et al., 2024d) is a machine-translated version of the ZsRE dataset in 12 languages: English, Czech, German, Dutch, Spanish, French, Portuguese, Russian, Thai, Turkish, Vietnamese, and Chinese. Translations were generated using Google Translate without human revision. Each sample is used to evaluate reliability, generality, locality, and portability.

biZsRE (Wang et al., 2024a) is a bilingual version of ZsRE translated into Chinese. Training data was translated using GPT-3.5-turbo, while test samples were translated via GPT-4 and then validated by native Chinese speakers.

Eva-KELLM (Wu et al., 2023) extends knowledge editing evaluation to full documents. Starting from the Counterfact dataset (Meng et al., 2023a), ChatGPT was used to generate documents corresponding to each edited fact. The dataset consists of 8,882 English documents and 6,930 Chinese documents.[4]

BMIKE-53 (Nie et al., 2025)[5] evaluates cross-lingual in-context knowledge editing across 53 languages. It unifies ZsRE, Counterfact, and WikiFactDiff datasets, translating them into 52 languages using GPT-4o. Translation quality was verified qualitatively via manual inspection and quantitatively via BLEU scores on back-translated samples.

CROLIN-MQUAKE (Khandelwal et al., 2024)[6] is a translated version of the multihop model editing datasets MQuAKE-CF and MQuAKE-T (Zhong et al., 2023). It includes translations into German, Spanish, Chinese, Russian, Hindi, Bengali, and Swahili. Translations

---

[4]Referred to as DocTer as of July 24, 2025
[5]Not publicly released as of May 15, 2025
[6]Not publicly released as of May 15, 2025

Dataset/Method	Single-hop	Multi-hop	Reliability	Generality	Locality	Portability
MLaKE	✓	✓	✗	✗	✗	✗
mZsRE	✗	✗	✓	✓	✓	✓
biZsRE	✗	✗	✓	✓	✓	✓
Eva-KELLM	✗	✗	✓	✓	✓	✓
BMIKE-53	✗	✗	✓	✓	✓	✓
CROLIN-MQUAKE	✓	✓	✗	✗	✗	✗
MQuAKE-AR	✓	✓	✗	✗	✗	✗
MQA-AEval	✓	✓	✗	✗	✗	✗
MKEB	✗	✗	✓	✓	✓	✓
Counterfact (Catalan)	✗	✗	✓	✓	✓	✓
CKnowEdit	✗	✗	✓	✓	✓	✓
BabelNet	✓	✓	✓	✓	✓	✓

Table 4: Comparison of datasets/methods across reasoning type and evaluation dimensions.

were reviewed by human experts and validated through back-translation BLEU scores.

MQuAKE-AR and MQA-AEval (Ali et al., 2025) include Arabic translations of MQuAKE-CF and MQuAKE-T. The translation process combined LLM-based scoring (via ChatGPT) with manual refinement of low-quality samples. MQA-AEval is a manually curated Arabic benchmark for evaluating multihop edits.

MKEB (Xie et al., 2024)[7] was constructed by querying Wikidata via SPARQL to retrieve factual triplets and chains. ChatGPT was used to generate corresponding questions, and translations into 12 languages (including English, Chinese, French, German, Japanese, Korean, and Portuguese) were performed via the Baidu API. No human revision was conducted.

Counterfact (Catalan) (Tamayo et al., 2024) is a translated version of the Counterfact dataset into Catalan. Translations were performed using the Aina English-Catalan translator and aligned using SimAlign (Jalili Sabet et al., 2020). Filters ensured sentence order consistency, and one round of human revision was conducted.

CKnowEdit (Fang et al., 2024) is a Chinese-centric benchmark designed to evaluate corrections of linguistic, factual, and logical errors. Unlike other datasets, it emphasizes language-specific and cultural relevance in Chinese.

BabelEdits (Green et al., 2025) is a knowledge editing benchmarking comprising 60 languages. It combines multilingual synsets from babel net with marker based translation to ensure entity (subject) translation quality. A comparison between the

dataset is shown in Table 4

---

[7]Not publicly available as of May 15, 2025